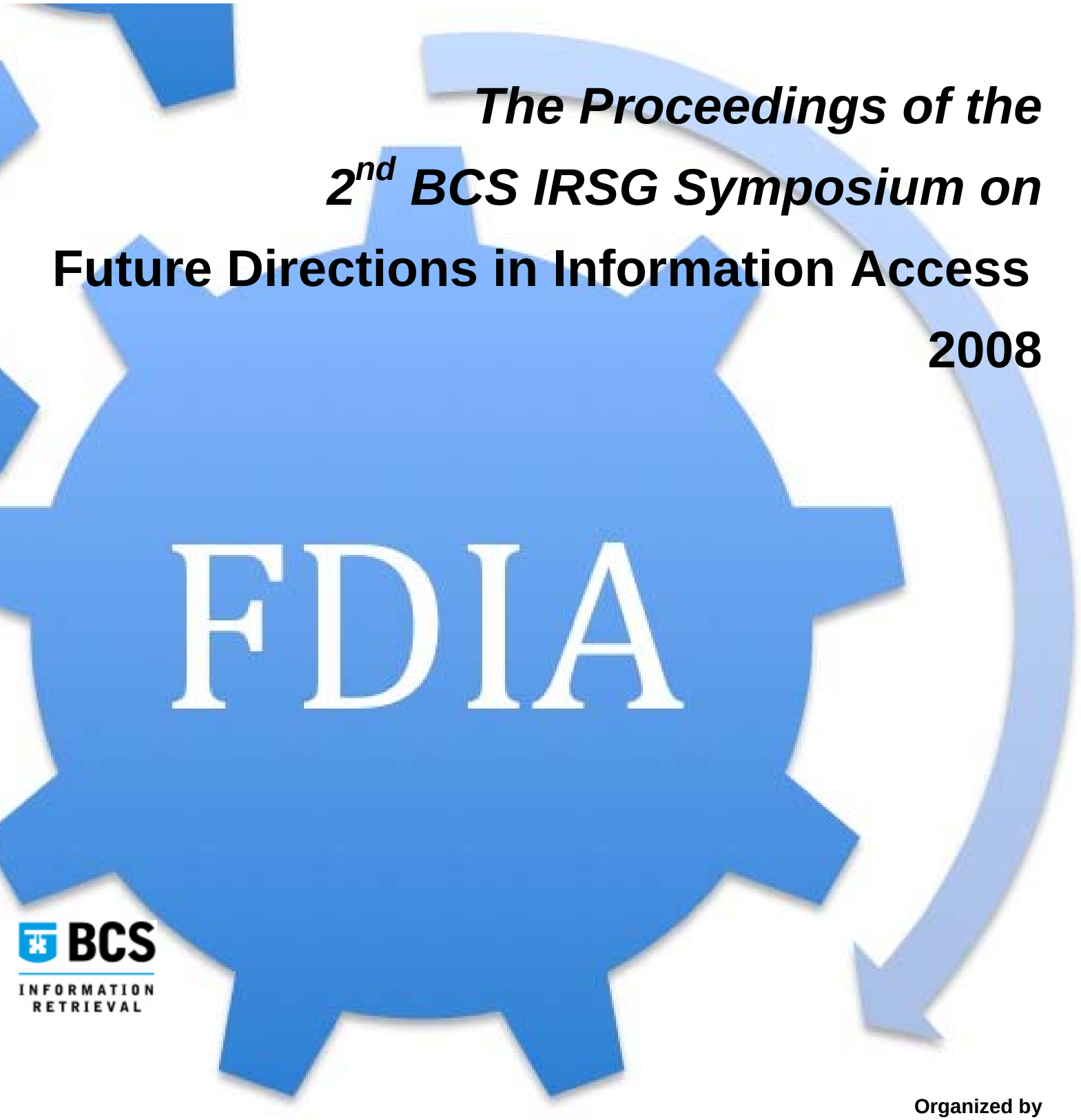
# *The Proceedings of the 2nd BCS IRSG Symposium on*

# Future Directions in Information Access 2008




BCS
INFORMATION RETRIEVAL

**Organized by**

**Leif Azzopardi**

**Andy MacFarlane**

**Murat Yakici**

**Ayse Goker**

**Published as part of the eWiC Series**

# Preface

**Future Directions in Information Access**

Last year the 1st BCS-IRSG Symposium on Future Directions in Information Access was established to provide a forum for early career researchers to present, share and discussion their research which is at a more formative or tentative stage.

**Symposium Aims**

The objectives of the Future Directions in Information Access (FDIA) are:

- To provide an accessible forum for early researchers (particularly PhD students, and researchers new to the field) to share and discuss their research.
- To create and foster formative and tentative research ideas.
- To encourage discussion and debate.

**Symposium Themes**

*Future directions*: to encourage research that focused on the possible paths and further work. Presenting the, what if scenarios, possible solutions, pilot studies, conceptual and theoretical work.

*Information Access*: to capture the broader ideas of information retrieval, storage and management to include interaction and usage.

**FDIA 2008**

These proceedings contain the papers presented at the Second BCS IRSG Symposium on Future Directions in Information Access (FDIA2008), held in London on the 22nd of September 2008 at the BCS London Office. FDIA 2008 was held in conjunction the BCS-IRSG Search Solutions held on the 23rd of September 2008. This year's symposium received eleven submissions of which ten were accepted for publication and presentation. Each submission was reviewed by two senior Information Retrieval researchers who were asked to provide detailed reviews and comments to help steer and guide the research presented. In order to facilitate more reviewing each reviewer was given one or two submissions to review. So we are very grateful to the members of the programme committee for their reviews of the submitted papers and we would like to thank them for their much appreciated effort.

Also, we would like to extend our thanks to the BCS for hosting the event, and in particular, Gemma Liddard, Rachel Browning and Mandy Bauer from BCS for their help and assistance in local organization.

**Leif Azzopardi**
**On Behalf of the Organizers**
**and the BCS-IRSG**

**Organizers**
Leif Azzopardi, University of Glasgow
Andy MacFarlane, City University
Murat Yakici, University of Glasgow
Ayse Goker, City University

**Programme Chair**
Leif Azzopardi, University of Glasgow

**Program Committee**
Alex Bailey, Google
Bettina Berendt, K.U. Leuven
Nick Craswell, Microsoft
Norbert Fuhr, University of Duisburg-Essen
Juan M. Fernandez-Luna, University of Granada
Ayse Goker, City University
Gareth Jones, Dublin City University
Gabriella Kazai, Microsoft Research Cambridge
Udo Kruschwitz, Essex University
David Losada, University of Santiago de Compostela
Monica Landoni, University of Strathclyde
Andy MacFarlane, City University
Massimo Melucci, University of Padua
Michael Oakes, University of Sunderland
Mark Sanderson, University of Sheffield
Dawei Song, Open University
Maarten de Rijke, University of Amsterdam
Stephen Robertson, Microsoft Research Cambridge
Stefan Rueger, Open University
Ian Ruthven, University of Strathclyde
Jun Wang, University College London
Murat Yakici, University of Glasgow


# Programme

## Monday, 22nd of September, 09:00 - 17:30

## Covent Gardens, BCS London Office

**09:00-9:30 Coffee on Arrival**

**09:30-10:00 Welcome Address**

**10:00-11:30 Session One: Context and Language in IR**



**11:30-12:00 Coffee Breakout Session**

**12:00-13:15 Session Two: Applications and Distributed Systems**



**13:15-14:00 Lunch Breakout Session**

**14:00- 15:30 Session Three: New Domains of IR**



**15:30-17:30 Session Four: Breakout Discussion Zone**: Meet, Mingle, Mix Session, with

- Poster presentations: comment, question, find out more
- White board analysis: explain, argue, enlighten
- Served with coffee and snacks, and later followed by a buffet and drinks.

BCS INFORMATION RETRIEVAL

# Modeling the Evolution of Context in Information Retrieval

Emanuele Di Buccio
Department of Information Engineering
University of Padova
Via Gradenigo 6/B, 35131 Padova, ITALY
*dibuccio@dei.unipd.it*

**Abstract**

**An Information Retrieval (IR) system ranks documents according to their predicted relevance to a formulated query. The prediction depends on the ranking algorithm adopted and on the assumptions about relevance underlying the algorithm. The main assumption is that there is one user, one information need for each query, one location where the user is, and no temporal dimension. But this assumption is unlikely: relevance is context-dependent. Exploiting the context in a way that does not require an high user effort may be effective in IR as suggested for example by Implicit Relevance Feedback techniques. The high number of factors to be considered by these techniques suggests the adoption of a theoretical framework which naturally incorporates multiple sources of evidence. Moreover, the information provided by the context might be a useful source of evidence in order to personalize the results returned to the user. Indeed, the information need arises and evolves in the present and past context of the user. Since the context changes in time, modeling the way in which the context evolves might contribute to achieve personalization.**
**Starting from some recent reconsiderations of the geometry underlying IR and their contribution to modeling context, in this paper some issues which will be the starting point for my PhD research activity are discussed.**

*Keywords: Information Retrieval, Personalization*

## 1. INTRODUCTION

Information Retrieval (IR) can be framed as a problem of evidence and prediction [19]. Indeed, the purpose of an IR system is to predict which documents are relevant to any information need of any user. What makes IR difficult is that, “like many other things in life, relevance is relative” [23]. In particular, there are many aspects which affect the information need of a user, and consequently the prediction of relevance. The information need depends on the user, the specific task the user is performing, the place and the time. In other words, relevance depends on context.

Since IR is context dependent, the ranking algorithm should efficiently and effectively exploit the information available from the evidence provided by context for predicting relevance. As different assumptions imply different retrieval models, the preliminary assumptions according to which the algorithm works are fundamental. Making these assumptions explicit is important for understanding which information is exploited in an IR model.

The main assumption is that context does not change in time. But this assumption is unlikely. Let consider, for instance, Relevance Feedback (RF) techniques. The idea underlying RF is that the first retrieval operation can be considered as a “initial query formulation” [21]. Some initially retrieved items are examined for relevance; then the automatic modification of the query can be performed by the system by using the feedback collected from the user — for instance adding keywords, selecting and marking documents. The modified query can be considered a “refinement” of the initial query. As shown in [9] in the event of the interpretation of RF in Vector Space Model (VSM), such technique can be interpreted as a form of query context change. The

effectiveness of RF suggests an investigation of the role of the evolution of the context in the retrieval process.

RF techniques point out not only that the information provided by the evolution of context should be considered, but also that such information should be involved in a way that does not require an high user effort. Indeed, even if RF has been shown to be effective, users are reluctant to provide explicit relevance information because they do not perceive it as being relevant to the achievement of their information goals. A possible solution is the adoption of techniques which are transparent to the user, that is “implicit”. Implicit Relevance Feedback (IRF) techniques [7] can use different contextual features collected during the interaction between the user and the system in order to suggest query expansion terms, retrieve new search results, or dynamically reorder existing results. One of the source of difficulties of this kind of techniques is the need of combining different sources of evidence, i.e. different contextual features. The complexity of these approaches is one of the reasons for investigating the problem in a principled way, that is for the adoption of a model-based development. One of the benefits of this approach is that all the assumptions are made explicit: this is crucial in modeling context in order to understand which elements of the context are actually considered, and above all in which way the relationship between such elements is modeled.

The research activity of my PhD will be mainly focused on Personalized Search, that is explicitly considering the aspects which affect the relevance judgments of the user in IR. In particular, the way in which the information provided by the context can be used to achieve personalization will be investigated. Since, at least initially, the problem of the personalization of the information access will be faced in a principled way, in Section 2, after that some of the previous works will be briefly reviewed, some recently proposed frameworks will be mentioned. The discussion about these frameworks will continue in Section 3, where some questions arisen during my preliminary investigations of the problem of personalized search and the study of such frameworks are reported. These questions might be a starting point for my research activity, whose main goal is the design and the implementation of a ranking algorithm for personalized search based on techniques, like the IRF’s, which do not require an high user effort.

## 2. PREVIOUS WORKS

Since IR is context-dependent, the development of an IR system should consider the information provided by the context. In order to develop a context-aware system, the factors involved and the relationship between such factors should be made explicit. Since the context provides information useful to predict relevance, why is such information not included in the design and the development of many IR system? In [20] the author provides possible answers to this question. A first reason for not considering the context is that, at least initially, such choice allowed the development of IR systems to be simplified. Another reason is that most IR systems are developed to satisfy the need of most of the people most of the time. Personalized Search starts from another hypothesis, that is information access should be personalized. The information provided by the context might be a useful source of evidence to achieve personalization.

The reason for investigating this issue is that previously proposed techniques, which involved some contextual information, were shown to be effective. Let consider a well-known technique, that is RF [22]. These techniques are based on the explicit participation of the user: the user assesses if the documents in an initially retrieved set are relevant, or can suggest or select a number of terms in order to refine its query — as already mentioned in Section 1, RF can be interpreted as a form of query context change. The high user effort required by this kind of techniques together with their effectiveness, suggest to find a way to preserve the benefits of RF and remove its burdens. A possible solution are the techniques based on IRF [7], which use information obtained by the interaction with the documents, for instance, to recommend query expansion terms or retrieve new document sets. The information collected by monitoring the interaction with the IR system can be useful to understand the way in which the information need evolves during the information seeking activities. Many of the IRF algorithms use only one contextual feature, for instance display time [26] or clickthrough data [5]. But, as suggested

in [26], ignoring other sources of IRF means that information about other aspects of the search context is lost. This loss may affect the contribution of the considered factors. A framework which allows multiple sources of evidence to be considered seems to be a necessary solution to exploit suitably the context and its contribution. A problem is to find a theoretical framework for describing the complexity of a system which exhibits contextual behavior and allows the contributions of these factors to be suitably combined. The heuristic-based development can be a possible strategy to address the problem. The latter approach is not negative in itself, but the theoretical framework is to be preferred because "all the assumptions are made explicit and can be reconsidered and refined independently of the particular retrieval algorithms" [24]. Let consider, for instance, the Probability Ranking Principle (PRP) [18]. The assumptions underlying this principle are explicit. Starting from a reconsideration of the classical PRP assumptions, in [2] a new theoretical framework for Interactive IR (IIR) is proposed. The basic rationale is modeling the evolution of the information need by considering that the user moves between situations. "A situation reflects the system state of the interactive search a user is performing" [2]. In each situation the user has a list of possible choices and a positive decision moves the user to a new situation.

The reconsideration and the extension of previously proposed solutions in order to include contextual factors, as in the case of the PRP, is a possible approach. A radical different approach is the one proposed in [25], whose subject is a complete reconsideration of the geometry underlying IR by suggesting the use of Hilbert's vector spaces.That work constitutes a first attempt of creating a novel and unified IR theory which will allow the emerging challenge of context-sensitive and multi-modal search to be addressed. The VSM is reconsidered also in [9], where the idea of using a basis of a vector space to represent context is proposed. This work discusses also how to model the evolution of the context by linear transformations from one basis to another. Since these works and the proposed geometry can be suitable approaches to address the mentioned issues, in Section 3.1 some of the ideas proposed in such works will be briefly reviewed.

## 3. CONSIDERATIONS AND POSSIBLE RESEARCH ISSUES

In this section some questions and some considerations based on my preliminary studies on the problem of Personalized Search are reported. They will be the starting point to investigate how to model the contribution of contextual features in the evolution of the information need.

### 3.1. Why exploiting the geometry of IR?

The starting point of my PhD research activity will be the reconsideration of the geometry underlying IR proposed in [25, 9] and the investigation of some tools provided by Quantum Mechanics (QM) to model the evolution of the information need.

A first reason to consider these approaches is that, as van Rijsbergen states in [25], "the geometry of the information space is significant and can be exploited to enhance retrieval". Let consider, for instance, the framework described in [9]. In that work the author assumes that a basis of a vector space is the construct to model context. The underlying interpretation is that "a vector is generated by a basis just as an information object is generated within a context" [14]. This interpretation makes possible to describe that an information object can be generated within different contexts: indeed, a vector can be generated by different basis. This framework was shown to be effective and general enough to include contextual features belonging to different contextual levels, particularly the interaction context [12] and the linguistic context[1] [14].
The latter work presents another contribution, that is the *probability of context*, which is the probability that an information object has been generated by a context. The author shows how the probability function proposed to compute the probability of context might discriminate between relevant and non-relevant documents. This function is a trace-based function inspired by the probability formulation in QM. The latter is one of the possible benefits of the adoption, proposed

[1] Here, the expression "linguistic context", is used to indicate the "users' context of meaning when they use a particular query term" [6].

in [25], of Hilbert vector spaces[2], one of the mathematical foundation of QM, as basis for a language to model IR. In particular, the mentioned trace-based function can be explained by using one of the results reported in [25], that is the adoption of the Gleason's Theorem [3] to connect Hilbert space and probability. Indeed, according to van Rijsbergen, is "the way in which the geometric structure is exploited to associate probability with measurements" that may be useful for IR. Giving the intuition underlying this view of measurement, might be useful to understand the motivation for reasoning in a more abstract way, that is using Hilbert's space instead of finite inner product vector space as in the VSM.

In QM a *state space* is associated to any isolated (physical) system, which in particular is a complex vector space with inner product [15]. A *state vector* — a unit vector in the system's state space — completely describes the system. Quantities to be measured, named *observables*, are self-adjoint linear operators[3]. The result of the measurement of an observable is one of the eigenvalues of the operator — or better of the matrix representing the operator — corresponding to the observable, "with a probability that depends on the geometry of the space" [25]. Van Rijsbergen states that, one of the interesting properties of this view of measurement, is that it is general and applicable also to infinite systems. He gives some hints about the possible reason for not limiting the investigation to finite systems, particularly with regard to the images. Since at the present time it is not clear which can be the best representation for images, we cannot exclude that complex numbers and infinite dimensionality might be useful to specify operations on this kind of information objects — for instance complex numbers are needed when Fourier transforms of signals are done.

### 3.2. Contextual Factors and Evolution of the Information Need

QM and Hilbert's spaces may be an intriguing formalism to describe the evolution of the information need because this framework is intrinsically based on the concept of state of a system and it is general and it may be applicable to several "non quantum" domains [16, 17]. Indeed, the framework does not specify which is the state space of a system and the state vector that describes the state of the considered system. There are several works which investigate the connection between QM and IR [25, 1, 10, 13].
Let consider the interpretation described in [13], where the user-information interaction is interpreted as a complex system. In particular, the user-information interaction is described by a state vector of the product space of the state space which describes the document, or the visit of the document, and the state space that refers to the user. An interesting issue might be investigating the evolution of the state of such system. Reaching this challenging objective requires to find an answer to different questions. Finding these answers can help design a ranking algorithm which is sensitive to the context of the user or of the document.

The first question is about the formalism. An in depth investigation of the Hilbert Spaces will be required to understand if this tool could be useful for the objective of my research activity and, in particular, to model the evolution of the information need. In [9] the author proposes to model the context change as a matrix transformation in the proposed geometrical framework. Until now, this technique has not been used to predict relevance. In QM the evolution of a closed quantum system is described by a unitary transformation, in particular by a unitary operator which depends only on the starting and the final time.

- How can this operator be defined?

The detection of the adequate contextual factors is a fundamental issue to be addressed because the observed data is mapped in the vector space basis which represents the context. Probably,

[2] An Hilbert space is a vector space with inner product that is *complete*. Let denote with $(\mathbf{x}, \mathbf{y})$ the inner product between the vectors $\mathbf{x}$ and $\mathbf{y}$, let $||\mathbf{x}|| = \sqrt{(\mathbf{x}, \mathbf{x})}$ be the norm induced by the inner product on the vector space. An inner product space is *complete* if every Cauchy sequences with respect to the defined norm is convergent. In this paper the Hilbert spaces considered are complex vector spaces with inner product.

[3] A linear operator $A$ on a vector space $V$ is an operator $A : V \to V$ which assigns to every vector $\mathbf{x}$ a vector $A\mathbf{x}$ and for which $\forall \mathbf{x}, \mathbf{y} \in V$ and for all scalars $\alpha$ and $\beta$, $A(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha A\mathbf{x} + \beta A\mathbf{y}$. The *adjoint* of $A$, denoted by $A^*$, is defined by $(A^*\mathbf{x}, \mathbf{y}) = (\mathbf{x}, A\mathbf{y})$. An operator is *self-adjoint* when $A^* = A$.

such contextual factors might be useful also in the definition of the operator describing the evolution of the context. As previously mentioned, the information about the interaction between the user and the system might be a useful source of evidence. But, since the information need is not only related to the user, but also to the task the user is performing, maybe only a subset of the available evidence should be used to refine the prediction of relevance.

- Which are the contextual factors that most strongly influence user behavior during information seeking-activities?

This issue is important also in the event of multimedia IR: the contextual factors could vary according to the medium of the query. Another issue is that a great number of factors can affect relevance and some of these factors may be hidden. The term “hidden” refers to the fact that some information is not only about the user or only about the documents, but characterizes the interaction between these two poles of IR.

- In which way this information can be discovered?

In [13] the author proposes a methodology for IRF which exploits the concept of entanglement. The state of a composite system is entangled if it cannot be written as a product of states of its component systems; this notion seems to fit the case of such properties that are not proper of the document or of the user, but characterize the interaction between the two. In [13] the Schmidt Decomposition Theorem [15] is adopted to determine if a state is entangled or not. Singular Value Decomposition (SVD), which the Schmidt Decomposition is based on, allows the “most influential” contextual factors to be detected. Another possible solution is proposed in [11], where a statistical framework based on Principal Component Analysis is utilized in order to discover the “hidden contextual factors”. Although these techniques provide a solution to the previous question, other existing techniques will be investigated in order to understand the differences in terms of contributions and efficiency. The mentioned techniques might help to find an answer also to the problem of the detection of the suitable contextual factors, a key issue when IRF techniques are adopted. Finding a solution to this problem will be important also when the theoretical results will be experimentally evaluated — some issues related to the experimentations are reported in Section 3.3. The previous mentioned decomposition techniques, as shown in [11, 13], might be a starting point to find an answer to the following challenging question:

- How can the different factors involved be suitably combined and used for the prediction of relevance?

Indeed, the purpose is understanding how the contextual information can be exploited to enhance retrieval. Once the adequate contextual factors are identified and the operator to model the evolution of context is defined, since the purpose is not the mere investigation of the existence of a transformation which describes the evolution of the system, but the final aim is the prediction of relevance, the main question will be:

- Does this operator help predict relevance?

In order to answer this question an in depth investigation of the behavior of the transformation should be done both at the theoretical and experimental level. This investigation might be useful to understand how the state of the system changes with time. Another question is if the QM can improve an approach based on the geometric interpretation of the context change proposed in [9].

### 3.3. Experimental Validation of the Theoretical Results

The evolution of the user-information interaction will not be investigated only at the theoretical level. The obtained theoretical results will be experimentally validated. But different issues raise because of the adoption of the experimental approach.

A first issue, already mentioned in Section 3.2, is the problem of the detection of the “most influential” contextual factors.

Another issue has to be addressed is the problem of the dataset. Indeed, information about the interaction between the user and the documents is required to test if the proposed approach can be useful to predict relevance and to achieve personalization. An example is the dataset utilized in [26], constituted by interaction logs obtained during a longitudinal user study of seven subjects' interaction behaviors over a period of fourteen weeks. The interaction logs store information like display time, number of keystrokes for scrolling a web page, if a page has been saved, bookmarked or printed. Part of my research activity will be focused on the implementation of a tool to collect this kind of information and which can be subsequently integrated with the functionalities of the ranking algorithm obtained by the theoretical investigation.
Another interesting but at the same time troublesome issue, is the choice of an appropriate measure of retrieval effectiveness. Precision and recall generally are based on binary relevance judgments. But, as stated in [4], the overwhelming number of documents the modern large retrieval environments return to the users, suggests the adoption of graded relevance judgments. The latter approach allows for developing IR techniques which identify highly relevant documents: indeed all documents are not of equal relevance to their user. Highly relevant documents should be identified and ranked first for presentation. A possible solution is a generalization of the measures of recall and precision or the development of novel measures based on graded relevance judgments. In [4] several novel measures are proposed, among which the Normalized Discounted Cumulative Gain (NDCG). The latter is devised to be able to handle useful score ranging in a non-binary scale and to make a better use of multi-level judgments than precision. The problem is that the definition of measures like precision and recall is based on prior human judgments. As a consequence, conclusions to such an experiment are very subjective as they are limited to the scope of the test collection and to the context. As pointed out in [1], a formal method for abstracting user behavior will allow to duplicate and verify experiments. The change of perspective due to the adoption of the QM mathematical framework, will affect also the field of evaluation, and provides an instrument to study a formal model which "approximates" the user. The problem of finding an appropriate measure of retrieval effectiveness might be an interesting issue to be addressed in future research.

## 4. CONCLUSIONS

This paper is focused on some issues concerning personalized search, that is explicitly considering the aspects which affect the relevance judgments of the user in IR. The main objective of this work is describing the motivation for focusing my PhD research activity on this problem, and in particular on a model-based development. There are several issues which are pointed out during my preliminary investigations. The main problem I am going to investigate is the evolution of the state of the system which models the user-information interaction and if the information obtained by the study of the evolution may be useful for the prediction of relevance and to achieve personalization. Some considerations derived from some recent investigations of the geometry of the information space have been reported. In particular, the motivation for starting from a geometry of IR has been explained, also with regard to the possible connection with QM. Although the relationship between QM and IR requires further investigation, some recent results, like the introduction of the probability of context, seem to indicate QM to be promising to model context in IR. Indeed, the quantum approach may be a suitable formalism when the system under study is sufficiently complex and exhibits contextual behavior [8]. Moreover, the tools QM provides in the fields of Statistics and Probability Theory, might be suitable tools to fit the uncertainty which intrinsically characterizes IR.

## 5. ACKNOWLEDGMENTS

The author is grateful to Massimo Melucci for the suggestions provided for this paper, and to the reviewers for the useful comments.

## REFERENCES

[1] Arafat, S. and van Rijsbergen, C.J. (2007) Quantum Theory and the Nature of Search. In *Proceedings of QI 2007*, Stanford, CA, USA, 26–28 March.

[2] Fuhr, N. (2008) A probability ranking principle for interactive information retrieval. *Information Retrieval*, **11**, 251–265.
[3] Hughes, R.I.G. (1989) *The Structure and Interpretation of Quantum Mechanics*. Harward University Press.
[4] Järvelin, K. and Kekäläinen, J. (2002) Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, **20**, 422–446.
[5] Joachims, T., Granka, L., Pan, B., Hembrooke, H. and Gay, G. (2005) Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of SIGIR'05*, Salvador, Brazil, 15–19 August, pp. 154–161, ACM Press. New York, NY, USA.
[6] Kelly, D. (2006) Measuring Online Information Seeking Context, Part 1: Background and Method. *Journal of the American Society for Information Science and Technology*, **57**, 1729–1739.
[7] Kelly, D. and Teevan, J. (2003) Implicit feedback for inferring user preference: a bibliography. *ACM SIGIR Forum*, **37**, 18–28.
[8] Kitto, K. (2008) Why quantum theory? In *Proceedings of QI 2008*, Oxford, UK, 26–28 March, pp. 11–18.
[9] Melucci, M. (2005) Context modeling and discovery using vector space bases. In *Proceedings of CIKM'05*, Bremen, Germany, 31 October – 5 November, pp. 808–815.
[10] Melucci, M. (2007) Exploring a mechanics for context aware information retrieval. In *Proceedings of QI 2007*, Stanford, CA, USA, 26–28 March.
[11] Melucci, M. and White, R.W. (2007) Discovering hidden contextual factors for implicit feedback. In *Proceedings of CIR'07*, Roskilde, Denmark, 20–21 August.
[12] Melucci, M. and White, R.W. (2007) Utilizing a geometry of context for enhanced implicit feedback. In *Proceedings of CIKM'07*, Lisbon, Portugal, 6–9 November, pp. 273–282.
[13] Melucci, M. (2008) Towards modeling implicit feedback with quantum entanglement. In *Proceedings of QI 2008*, Oxford, UK, 26–28 March, pp. 154–159.
[14] Melucci, M. (2008) A basis for Information Retrieval in Context. *ACM Transactions on Information Systems*, **26**, 1–41.
[15] Nielsen, M.A. and Chuang, I.L. (2000) *Quantum Computation and Quantum Information*. Cambridge University Press, UK.
[16] Quantum Interaction 2007. http://ir.dcs.gla.ac.uk/qi2007.
[17] Quantum Interaction 2008. http://ir.dcs.gla.ac.uk/qi2008.
[18] Robertson, S.E. (1977) The Probability Ranking Principle in Information Retrieval. *Journal of Documentation*, **33**, 294–304.
[19] Robertson, S.E., Maron, M.E. and Cooper, W.S. (1982) Probability of relevance: a unification of two competing models for document retrieval. *Information Technology: Research and Development*, **1**, 1–21.
[20] Ruthven, I. (2004) "and this set of words represents the user's context..." In *Proceedings of the SIGIR 2004 Workshop on Information Retrieval in Context*, New York, NY, USA. ACM Press.
[21] Salton, G. and Buckley, C. (1990) Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, **41**, 288–297.
[22] Salton, G. and McGill, M.J. (1983) *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, NY.
[23] Saracevich, T. (1975) Relevance: A Review of and a Framework for the Thinking on the Notion in Information Science. *Journal of the American Society for Information Science*, **26**, 321–343.
[24] Teevan, J. and Karger, D.R. (2003) Empirical development of an exponential probabilistic model for text retrieval: using textual analysis to build a better model. In *Proceedings of SIGIR'03*, Toronto, Canada, 28 July – 1 August, pp. 18–25, New York, NY, USA. ACM Press.
[25] van Rijsbergen, C.J. (2004) *The Geometry of Information Retrieval*. Cambridge University Press, UK.
[26] White, R.W. and Kelly, D. (2006) A Study on the Effects of Personalization and Task Information on Implicit Feedback Performance. In *Proceedings of CIKM'06*, Arlington, Virginia, USA, November 5–11, pp. 297–306.

# Integrating Memory Context into Personal Information Re-finding

Yi Chen Gareth J F, Jones
Centre for Digital Video Processing, School of Computing, Dublin City University, Dublin 9, Ireland
*{ychen,gjones}@computing.dcu.ie*

**Abstract**

**Personal information archives are emerging as a new challenge for information retrieval (IR) techniques. The user's memory plays a greater role in retrieval from person archives than from other more traditional types of information collection (e.g. the Web), due to the large overlap of its content and individual human memory of the captured material. This paper presents a new analysis on IR of personal archives from a cognitive perspective. Some existing work on personal information management (PIM) has begun to employ human memory features into their IR systems. In our work we seek to go further, we assume that for IR in PIM system terms can be weighted not only by traditional IR methods, but also taking the user's recall reliability into account. We aim to develop algorithms that combine factors from both the system side and the user side to achieve more effective searching. In this paper, we discuss possible applications of human memory theories for this algorithm, and present results from a pilot study and a proposed model of data structure for the HDMs achieves.**

*Keywords: Personal Information Management, Human Digital Memories, Re-finding, Cognitive Memory Model*

## 1. INSTRUCTION

Development of hardware recording devices together with associated recording software, and reductions in the cost of in digital storage is now allowing vast digital archives of personal life experiences to be captured. These personal archives, which we call Human Digital Memories (HDMs), can contain various types of data created or accessed by the individual. While most of present 'life logging' projects are still confined to recording users' activities in the 'digital world', mainly concerning user's interactions with electronic files (e.g. documents, emails, images, videos, etc.) that they have accessed on their computers [e.g. 2, 3], another strand of work, which is aimed at recording the real life, is beginning to develop. Research in this area usually involves wearing audio or video capturing devices to track the user's behaviour in a laboratory environment [e.g., 4, 6]. Many potential benefits have been put forward for such systems. According to Sparck Jones [5], these archives can be used in following ways: storing the past information for a person as an 'Deposit'; amplifying people's memory of events as a 'super me', by providing linked relevant information; showing one's past life (or certain aspect of it) to other people; sharing memory of certain information among different people. By far the most common proposition for real life logging is to provide support for people's memory about their past by presenting them with data captured during their daily activities [e.g, 1, 6]. However, for any of the above applications, it is essential for an individual to be able to locate and retrieve the desired items from them. We aim to explore an efficient way of information retrieval (IR) for this new type of data collection.

We structure this paper as follows: Section 2 introduces some related work on integrating context data in IR of life logs, and basic notions of human memory with an associated memory model, which deals with how information is encoded and retrieved with contextual cues. Based on his model, we explain how memory works when people perform information re-finding tasks. Section 3 presents results from our pilot study on memory of photos; Section 4 gives a brief description of our present data collection work and presents our proposed model of linking and weighting the collected data in retrieval.

## 2. BACKGROUND

Present life logging data usually involves multimedia data such as audio, video or static images. Due to the huge amount of data, it would be a time consuming and tedious to look for a certain frame of scene from one year's video recordings, merely by browsing. For the same reason, it is almost impossible for the users to annotate these data sources manually. Lack of textual content in these types of data makes them difficult to retrieve by traditional content-based text IR methods. Current IR techniques for audio and images collections are based on content analysis of low level features. The lower quality of HDM data may significantly reduce the accuracy of automatic semantic annotation with content-based methods. Indexing these types of information with their embedded timestamps may be a solution. Yet, just like other textual types of data, traditional query based searching in IR systems relies too much on the user's ability to recall accurate details about the searching target, such as the key words, titles, and the exact time.

## 2.1 RELATED WORK

Existing studies have been trying to cope with the problem of indexing features by allowing manual annotation and then searching files with well remembered types of features such as episodic context information [12]. Hori et. al.[17] have tried to log people's life with wearable camcorders. They used simultaneously captured context information to help index and retrieval. Apart form the microphone and camera embedded in the camcorder, they also had a GPS device to provide address keys, latitudes and longitudes; Biometric devices such as acceleration sensor, and a gyro sensor and a brain wave analyzer were employed in their system to capture the wearer's motion and mental status information. Face detection technology was also used to indicate the people shown in the captured video. With this context information embedded in the life log video, their system allowed the user to locate and retrieve required episodes of video with information about their corresponding experiences, such as their emotion, location, action, weather and other people shown up.

Other example systems include Mylifebits[7] and MediAssist[14]. In Mylifebits, contextual information such as location, people, and date are used, they do not only assist retrieval, but also link events by these data [7]. MediAssist uses extended types of both content and contextual information in a photo retrieval system, and allows more flexibility, for example, instead of limiting the date time to accurate numbers, which the user may not necessarily knew even at the time of the occurrence of the event, it enable them to roughly decide the range or period, which is more likely to be perceived and therefore remembered from the time of capture [14].

Significantly personal archives are different from other more traditional information collections in that they comprise information that is or used to be in the owner's memory. Therefore, they possess more potential for interaction with the features recalled from the owner's memory. We assume that if well remembered information can satisfy searching needs, it may to some extent relieve the burden of human memory at the time of re-finding, and make the IR (based on the information that can be recalled) more reliable. Combined with the presentation of related clusters and iterative rounds of searching outlined above, we believe that this can form the basis for improved search techniques suitable for the HDM domain. As very little computer-human interaction literature explores the link between life-logging technologies and human memory, we aim to put more effort in this aspect by exploring the possible application of human memory from cognitive theories or models into technologies for IR from HDMs. To have a better idea of how to utilize human memory for more efficient and less effortful searching, we first look into the information processing of human memory.

## 2.2 Human Memory Information Process: Basic Concepts

*2.2.1 Encoding and working memory*

Encoding can be defined as a set of operations that people use to code incoming stimuli. These processes modify and organize the arriving data by associating it with information in their current memory. Encoding involves sensory memory, working memory, as well as long term memory. Sensory memory, which is also called sensory registration, stores massive amount of raw data from physical stimuli features very briefly (<1 second) before processing them. This kind of memory decays very quickly and is replaced immediately when new sensory information is registered. Working memory holds limited information (typically limited to 7±2 items) for a longer time (usually no more than a few minutes) while processing them [1].

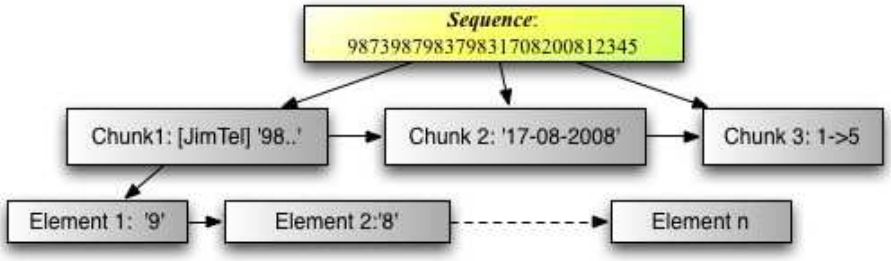



**FIGURE 1:** Chunking Structure

Due to the limited capacity and durance of this short term storage, some memory strategies are either intentionally or unconsciously employed to make maximum use of each processing stage. These strategies are particularly important for explicit intentional learning. Chucking is one strategy most frequently observed in learning experiments. It breaks big information sets into small pieces randomly, but usually base on sequential proximity [3]. Strategic chunking applied in learning enables short term storage holding 7±2 chunks of items, which can be expanded to smaller units [1]. For example, if you were asked to remember a list of random digits, 98739879837983170820081234 5, it would be very difficult to hold them all in mind to repeat. However, if they were broken into pieces, for instance, 9873987983 happens to be a phone number you remembers well, 17082008 can be considered as a date17-08-2008, then user can chunk them into 9873987983- 17-08-2008-12345; The chunked sequence is then much easier to hold in short term memory. The strategy behind this chunking might be to extract the lower level elements from long term memory where the chain of these lower level elements resides. The lower level items are of strong association, thus requiring less cognitive or attentional resources to retrieve them one by

one. This process schema may be able to inspire some algorithms in IR, which for instance, only processes the higher level nodes (chunks).

*2.2.2 Storage and retention*

Storage refers to the process of placing the coded information into memory system for long term storage, similar to saving a file on the hard drive. Representation is the format in which the information is stored in memory, e.g. spatial, auditory, and semantic [1]. One of the most important issues in memory storage is the retention of memory, which means how well the information is stored. Among many studies investigating this issue, the basic two most widely consented factors determining retention are: time lapse since encoding and frequency of repetition. It also depends on the initial strength determined by the encoding quality at learning [18]. Information processing theories have suggested that human memory exists in an associated network where the nodes of memory are bi-directionally linked; the stronger the link means the easier to evoke the associated information [19]. During each repetition, the item is re-encoded, thus the memory of one item may be bonded (linked) to various contexts. While the links between the item and most of the context information fade overtime, some association may be reinforced due to repetition. Also, as some information (context) is more readily associated with the item (target at the retrieving task), possibly due to their pre-existed relationship, the link between this context information and the item is likely to encoded with greater strength and retained for longer. For this reason, we can expect an uneven distribution and dynamic change of memory about the items and their related context.

*2.2.3 Retrieval and forgetting*

Retrieval is the process of recollecting information from long-term storage, and presenting the output [18], which can either be detailed information of the retrieving target or a judgement of whether the target exists in one's memory. These two types of outputs are referred to as remembered and known. The remember/know paradigm is widely used in studying the human memory, corresponding to recall and recognition tasks respectively.

According to psychology literature, forgetting (of the past) is usually caused by the inability to retrieve the item from memory rather than the lost or damage of storage. That is, the inability of tracing back to where that piece of memory is stored [1]. It explains why cued recall tests (prompted with related information) usually results in better recall performance than free recall tests (with no provided cues nor subject to any specific order). Thus, we assume that a user can have better recollection of required information for searching (queries), if they are presented with relevant information as cues. As for free recall, most of the ideas that pop out are in fact either cued by the previous recalled piece of memory (thoughts) or triggered by external environment, which possesses certain elements that are associated with the piece of information in long term memory. Clustering is usually observed in free recall tasks. It can be considered as a more advanced version of chunking in retrieving, as it groups information according to some higher-level criterion, such as cross-modality similarity [16]. Forgetting has also been argued as an mechanism of filtering out unwanted memory, that means it is possible that people forget things because they do not think they want to remember them (i.e. unimportant, or cause too much pain remembering that) [1]. Yet, it is also possible they might want to use this previously unwanted information, but are unable to retrieve it due to forgetting. In this case, an HDM can potentially shows its advantage as a 'repository'. However, it also faces the same problem as forgetting in human memory of being unable to locate the information.

*2.2.4 Context Factors that Influence Memory Retrieval*

Retrieval and the popping out of associated (or clustered) ideas are triggered by the interaction of external information and internal context [20]. Context is a vague concept. In the physical world, it usually refers to the external physical environment, including temporally and specially surrounding information, which is assumed to be encoded together, associating with each other, and acting as cues at retrieval. These types of related information present at the time of encoding or retrieval, are called *external context*, while the *internal context* is the activated information in human memory. When new information comes in, it broadcasts to the stored memory and the pre-existing nodes which are active enough (above the 'to be perceived' threshold), these nodes react to the broadcast if they can be associated with the incoming information according the clustering rules. The threshold is determined by the effort or energy at the time of broadcasting. Those high priority links with best matched nodes pop out and construct the internal context which interacts with the input and leads to the searching target. It resides in short term storage waiting to associate or to reinforce association with the input information [21].

### 2.3 Human Memory in Information Re-finding Tasks

When an individual is motivated to re-find some information from their previously accessed information (e.g. files on their desktop), the first recalled pieces of information of the item will trigger clusters of memory which they belong to. These clusters of information form the internal context which contains both the information of the external context in the events when encoding the item, and those associated with the item based on (cross modality) similarity. There two general types of memory cluster, either episodic memory of the related information of the events during which the item was encoded, or semantic memory where the item conceptually resides in (e.g. information of the same the category). With the memory searching criteria provided by the system interface, the individual could decide which route or cluster to look into in order to retrieve the corresponding memory. For example, to search for (or re-find) previously accessed data, users need to look into their own memory to find

information which is needed to perform the searching task. For example, the Microsoft Windows desktop search (WDS) allows the user to search by file name, path, file size, etc. while searching for his or her previous accessed files, the user needs to recall the information that the above fields require, in order to be able to perform the search. The recalled information will then be transformed into the form of a query to perform the IR task in the search engine. The matched results will be returned to the user to judge and decide whether one of the items is the requested target. When these results are presented, they act as an external input, which will modify the user's internal context. This change of context (usually new information is added) means that there is a possibility of recalling more related information, which does not belong to the clusters in the previous internal context. The newly recalled information may therefore lead to another round of searching. Admittedly, the user's memory may make mistakes, going to the wrong route (and misleading queries) or get a false alarm in recognition of suggested information. In these cases, an automatic estimation of the memory retrieval reliability would be a great help.

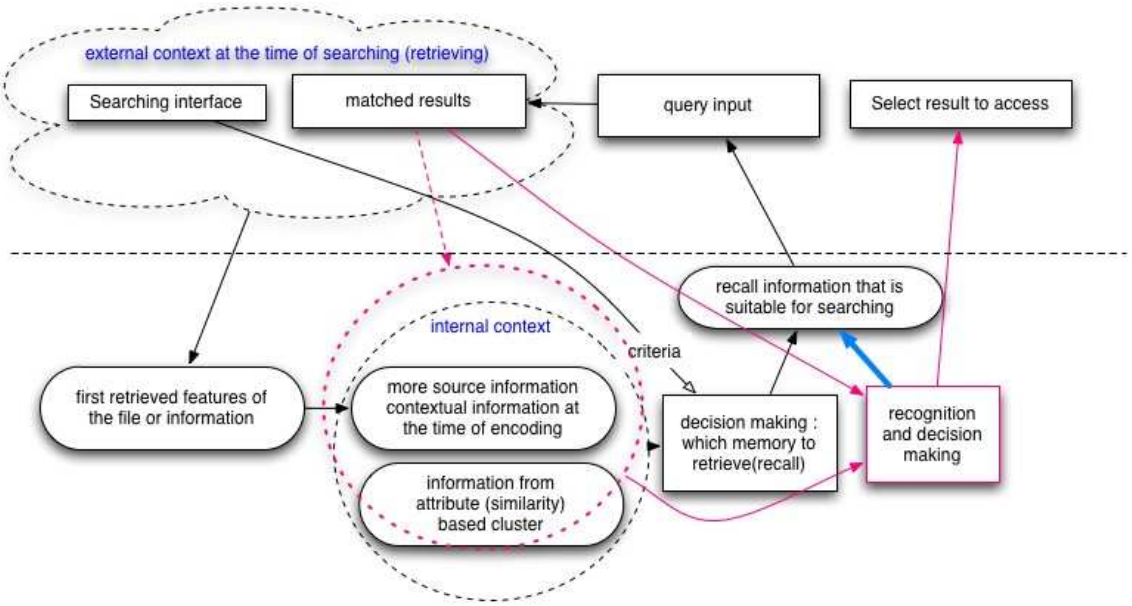



**FIGURE 1:** How human memory works during Information Re-finding

### 2.4 Related work on memory of personal data

A previous study tested people's memory about several types of content and context information of their own photos [12]. The participants were required to free recall 3 tasks followed with an implicit evaluation of the features of their interface, which offered several categories of context fields. Their study found that outdoor/indoor classification, location and people were all well recalled and proved to be useful cues for retrieval. Location and time derived context such as weather, daylight status and season were also proved to be useful. In particular, local time and daylight status seem to be stronger cues than the date/time of the photo, or even actual year if the photo was taken long ago. Finally, people also seem to remember photo colours, but its role as a retrieval cue was not tested.

Further studies explored memory on extended types of data from a small HDM, with a concern of decay of memory over time [6][11]. Memory of context information (involved in this data collection) showed a generally consistent pattern with [12]. Comparing with the participant's recall performances 6 month ago, we saw a varied degree of memory fade for different context information sources for the searching targets, e.g. semantic memory about textual information may decay faster than the episodic contextual information; also different behaviour of accessing the files influences and affects the memory recollection, e.g., self-generated data were generally better recalled than passively presented information. We also noticed that the keywords (content) which the participant selected to describe the documents 6 months ago were not necessarily the same as she recalled 6 months later. This finding is congruent with the memory model mentioned above, in that, as people's knowledge changes, their internal memory context may became difference from the target time (6 month ago) this item was encoded, therefore, the first few key words that the file links to changed. Thus, for the files whose retrieval is mainly based on metadata (e.g keywords), the mismatch between recalled features and the annotation might lead to a decline in their reliability for use in retrieval.

Traditional ways of automatically extracting keywords or summary content are based on statistics, e.g. term frequency (tf) and inverse document frequency (Idf). However, according to above findings, the most significant items as indicated by *tf-idf* weights may not necessarily be the ones that users are most likely to remember accurately for an item, although the frequent appearance of certain terms may to some extent improve the user's memory about them. A possible application of this finding is to modify the tf-idf score with recency and frequency components. For example, we may assume that the more recently frequently encountered terms are more important, as they are more likely to be recalled and therefore used it as queries.

According to various reasons mentioned above, static weighting of the metadata may be insufficient to achieve the best possible IR effectiveness. Besides, personal differences, e.g. different life style (for example, some people have most of their activity at the same place, thus location may not be distinctive for retrieval, while some people are frequent travellers, the locations may be a more distinctive and strongly associated context information) may also benefit from a more dynamic way of weighting the information. We believe that dynamic annotation or weighting of the annotations (or metadata) according to likelihood of being recalled correctly may be a solution for this issue. One of our goals is to explore how to use the above memory model to estimate the memory of individual metadata based on the data we capture. To estimate the recall possibility, an algorithm is needed to calculate and update the memory status of the data owner. According to the associated memory model, the ease or likelihood of memory being evoked and retrieved depends on the strength of links (either direct or indirect) to the provided cues. One essential step is to establish these links.

## 3. PILOT STUDY

We conducted a further new pilot study on memory of photos from the participants' own photo collection, to explore the types of association human memory uses. The reason for using photos is that they have embedded contextual information in themselves, which means no specific data collection for the experiment is needed. The study described in [12] has already provided very useful and detailed data on well remembered features for photo collections such as location, indoor/out door, people, and weather. In this study, we want to further explore the reasons and factors that influence memory and retrieval of photos.

### 3.1 Methods

*Subjects:* Three graduate students (all around 25 years old) participated in this study, two of them were interested in travelling and photographing, and the other one had some knowledge of multimedia IR.

*Material:* An electronic questionnaire was used to test their recall performance. It included one field for content, one for the photo's location (e.g. folder name and path on their computer), and other fields for contextual information, including 'your location', 'year', 'season', 'month', 'day' 'date', 'period of day', 'weather', 'people around'. This questionnaire also allowed the user to hide the content that they want to keep as private. Different from [6][11], the participant did not intend to collect data for re-finding task, which means that it was collected in a more natural setting.

*Procedures:* Instructions were given for each field before they started. They were asked to free recall any 20 of their photos from any period of their lifetime, and to input the photo's content and context into corresponding fields in the questionnaire, as soon as the photo popped into their mind. The participants were also asked to choose whether they took the photos themselves, which was assumed to indicate self-involvement, as according to our previous study, self-generated files were generally better remembered than passively present files.

A post-test face-to-face interview was also conducted to investigate why and how they recalled these particular photos, and if there are certain associations of neighbouring photos. This step aimed to explore why some photos are better remembered, and how the external context at the time of retrieval (e.g. previous photo) triggers memory activities. For example, one may recall a photo which is not so meaningful, but may be somehow related to the previously recalled one. Finally, the participants were asked to re-find these photos and check their recall results.

### 3.2 Results and Discussion

The results are generally consistent with the findings in [12]. Experience related information such as the participants geographical location (98%), weather (or light status) (90%), the season (89%) and period of day (65%) were well recalled, while the exact day/date (12%) is seldom remembered unless the number is particularly bounded to the event, e.g. one participant only went to a certain place and met certain people on Fridays, thus if he remembered the photo was taken on those occasions, 'Friday' can be deduced. 'Months' are well remembered by two participants for the photos taken during their travels (82%), but not by the participant who did not have much travel experience (32%). 'Years' are usually well remembered (96%). unless the event was remote, e.g. more than 5 years. 'People around' are also well recalled, though it cannot be checked only from the photos. 'Locations' (path) of the photos also had sound recall, possibly because that the participants organize their collections very well. Yet due to the considerable number of photos in each folder, the participants reported a searching tool might be of great help. However, the self-generating effect (photo taken by the participants themselves) did not show the same advantage as it was in our previous pilot study. The result may have some ceiling effect due to that the first recalled photos might happen to be what the participants remembered well. This result suggests that in our future studies certain criteria of searching which increase the difficulty of recalling should be employed.

The reported reasons for recalling photos generally fell into five categories (or combinations of these categories): interesting (46%), novelty or impressive (31%), frequently seen (18%), recently viewed (15%), and 'no particular reason, but it popped into my mind' (6%). The last category suggested a possibility that previous ones triggered the recall of these photos. This kind of trigger was also found possible in other neighbouring photos. This finding gives an example of the memory clustering mechanism at retrieval. The association found in this study included: things in the same event (e.g. in the same trip), similar occasions (e.g. gathering together with certain group of people, or

during festival celebrations), same location or similar type of location (e.g. train stations, parks, indoor). This implies a potential need to link files based on such kinds of information, and sheds light on interface design for presentation of results, e.g. to group the results by user preferred clustering approaches. Further study will be needed to explore other possible association or clustering approaches employed by human memory at retrieval time, with more data types.

When the participants looked back into their photo collections, there were a considerable number of photos that they could not recognize, and thus have no recollection of any information referring to these specific photos. This means these photos will not be searched. However, such photos might match some of the primary searching purpose, though they did not have these specific items in mind while performing the searching (re-finding) task. It would also be helpful to present these photos to them when what they actually want is a cluster of results, not specific files. Alternatively, the user may not wish to see the photos as all, and it would be interesting to explore means of suppressing the retrieval of items that are unlikely to be of interest to them at that moment. Again, the linking of files based on certain criteria, which the human memory uses to cluster information at retrieval, is desired.

## 4. CURRENT AND FUTURE WORK

### 4.1 Data Collection

In order to do further research on extended types of data over a much longer period, we have started a long term data collection exercise. This will be gathered from across one-year period with four participants. We are collecting computer activity, and other related data from their daily life, for example:

- *Computer activities* are logged by desktop software, Microsoft Digital Memories1 (DM) and S'life, every time an application comes to the foreground. The full textual content of the documents, WebPages, and other applications, as well as the file name and path, etc. are be recorded.

- *SenseCam image*: The SenseCam [8] is a wearable camera, which passively takes photos with its fisheye lens, and store in 640x480 JPG format. It can be triggered by scheduled time (e.g. every 20s) or change of environment such as a detection of people, change of luminance. It takes about 3000 photo per day from one person's daily life.

- *Bluetooth* is used to detect the surrounding Bluetooth devices, which can indicate the corresponding people (e.g. who have the Bluetooth on in their mobile photos) and objects (e.g. computers with Bluetooth on) [3]. Content based technologies such as face detection may also be a supplement, but not the main approach to detect people, due to the low quality of SenseCam images. It is also because that the wearer may not necessarily know the content of the SenseCam images, e.g. who were in the photos, since they may not want to review the images captured everyday. Also, as they may remember their experience context, such as who was present in an event, people who were nearby may not be captured in their SenseCam images, but may be a useful retrieval cue.

- *Geographical location*: we use GPS function on Nokia N95 mobile phones to record the individual's location [3].

- *Biometric devices*: wearable biometric devices such as *heart monitors* and *BodyMedia SenseWear armband* are used record the participants' physical conditions, which can to some extent, indicate the wearer's motion, emotional status and arousal level [10].

- Similar to many other life logging projects, one major concern of our data collection is the privacy[4], not only of the participants whose life is 'recorded ', but also of the third party individuals who may be somehow involved in the recording. For example, people shown up in the photos taken by Sensecam, emails or messages sent to the participant involving the individual's private information, etc. Consent forms were signed by the participants, and they are allowed to delete the data they are unwilling to show the developers.

### 4.2 HDMs Model: Structuring Raw Data

We aim to develop a memory model for data in the HDMs which can estimate the strength of memory features. In this model, we define following objects:

- Item: an attribute or related context information of the file, e.g. the title, the location.
- Link: There are difference types of links, as they are created according the rules of clustering memory. At this stage, we assume all the links to be bi-directional.

#### *4.2.1 Weighting of Links*

[1] http://research.microsoft.com/erp/memex/presentations/MSR Digital Memories 2006 Jim Gemmell Software.ppt

This model aims to link items and estimate the corresponding memory retrieval strength, which means how likely an item on one end of the link can be retrieved when cued by that on the other end. Thus, instead of weighting individual items, we weight the links. Therefore, for example, the stronger the link between one attribute and a file (a combination of several links of its metadata), the higher score this attribute will get for this file. Based on the theories of memory and learning [18], we propose three main factors: time lapse, which means the more recent encoded information, is more likely to be recalled [1]; frequency of repetition, that is the more often one comes across such information, the more likely it is remembered [18], it can also be applied to the term frequency in a document; encoding quality, which refers to how well the information is encoded which depends on several factors, such as arousal level, easiness of association, and distinctiveness [18]. For example, it is usually easier to remember things when you are fully awake than when you are very sleepy. And it is easier to remember the content of the paper if it is about your daily life, than if it is about advanced mathematics, assuming that you are not majoring in maths. The factor of distinctiveness at encoding comes from the notion of inhibition. That is, when you have seen something for several times, you are less likely to allocate attention to it, which means this information will be less well encoded. It is similar to the inverse document frequency in traditional IR.

*4.2.2 Processing of Data*

The full details of files and digital images, as well as the corresponding contextual data (location, emotional status, etc.) at the time of accessing or creating them will be recorded automatically. When the files and the above data are uploaded to the server, we also 'encode' them into a structured database. For example, all the context information as well other metadata such as keywords will be linked to the files, and they may link to each other based on memory clustering rules, which are yet to be explored.

**TABLE1:** table 'item' (Exmaple database table)

| itemID | Itemtype | path | Content |
|---|---|---|---|
| 00240000 | 'location' | | 'My kitchen' |
| 00240001 | 'people' | | 'Alice' |
| 00240002 | 'content' | C:\sensecam\img172.jpg | 'Spaghetti ' |
| …. | …. | …. | …. |
| 00250006 | 'location' | | [Lat: 52.480747 ing: 1.891784] |
| 00250007 | 'people' | | 'Alice' |
| 00250008 | 'file' | C:\sensecam\img234.jpg | |

**TABLE2:** table link' (Example database table)

| linkID | End1ID | End2ID | LinkType | lastTIMEaccess | Weight |
|---|---|---|---|---|---|
| 00003400 | 00240001 | 00240000 | 'Episodic'(attributes belong to the same event) | 12/06/2007 11:20 | 12 |
| … | … | … | … | … | … |
| 00007421 | 00250008 | 00250007 | 'fileAttribute' | 20/03/2008 15:21 | 22 |
| 00007422 | 00250008 | 00240001 | Filebyattribute | 20/03/2008 15:21 | 5 |

* Note: above table are only samples, the content of the fields, especially for Linktype and itemtype, is yet to be decided.

The weight of links, after getting an initial value at first time encoding, can be updated every time when any link in the network is changed. For example, when item 'A' (new item) is linked to 'B' (pre-existed nodes), other links with B will be weakened. Also, memory decay due to time lapse should be updated as often as possible. In this model, an update will be triggered by every encoding, time schedule, or manually.

**4.2 Problems of Realizing the Model**

Although the raw data will also be stored in the form of independent files and full-text indexing in Digital Memory and S'life, the atomic unit of information in this model will not be files, but the pieces of information, e.g. keywords, phrases, attributes or other metadata of the files. We will try to explore the digital elements of content that correspond to output of memory representation or recall. While there are many theories in memory studies of clustering information in retrieval, we still need to explore suitable approaches of associating information in HDMs.

**4.3 Other Possible Applications (Interface)**

According to the context congruent retrieval point of view [1], the better match between the context at the time of encoding and the time of retrieval, the more likely the item can be retrieved; also, the same modality of cues may have a better chance of triggering each other. Although we cannot always change the searching interface according to the target before searching, the way of presenting results can be varied. For example, the result can be presented as representative cluster nodes in a way that people cluster information in memory retrieval. Besides, suggestive interface presentation of related information which may trigger recall of potential queries may also be a solution. For example, information associated with the query within the same events or categories. With this model, which structures data by estimating the owner's memory status, we believe many other interesting applications may also be developed.

ACKNOWLEDGEMENTS

This work is funded by grant CMS023 under the Science Foundation Ireland Research Frontiers Programme 2006.

REFERENCES

[1] A. D. Baddeley. *Your memory : a user's guide*. Carlton, London, UK, 2004.
[2] G. Bell. A personal digital store. *Commun. ACM*, 44(1):86–91, 2001.
[3] D. Byrne, B. Lavelle, A. Doherty, G. Jones, and A. F. Smeaton. Using bluetooth and GPS metadata to measure event similarity in sensecam images. In *IMAI'07 - 5th International Conference on Intelligent Multimedia and Ambient Intelligence*, pages 1454–1460, 2007.
[4] W. C. Cheng, L. Golubchik, and D. G. Kay. Total recall: are privacy changes inevitable? In *CARPE'04: Proceedings of the 1st ACM workshop on Continuous archival and retrieval of personal experiences*, pages 86–92, New York, NY, USA, 2004. ACM Press.
[5] D. Elsweiler, I. Ruthven, and C. Jones. Towards memory supporting personal information management tools. *J. Am. Soc. Inf. Sci. Technol.*, 58(7):924–946, 2007.
[6] M. Fuller, L. Kelly, and G. Jones. Applying contextual memory cues for retrieval from personal information archives. In *PIM 2008 - Proceedings of Personal Information Management, Workshop at CHI 2008*, 2008.
[7] J. Gemmell, G. Bell, and R. Lueder. Mylifebits: a personal database for everything. *Commun. ACM*, 49(1):88–95, 2006.
[8] S. Hodges and Williams. *SenseCam: A Retrospective Memory Aid*. 2006.
[9] T. Hori and K. Aizawa. Context-based video retrieval system for the life-log applications. In *MIR '03: Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, pages 31–38, New York, NY, USA, 2003. ACM.
[10] L. Kelly. Context and linking in retrieval from personal digital archives. In *SIGIR 2008 -Doctoral Consortium, 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2008.
[11] L. Kelly, Y. Chen, M. Fuller, and G. Jones. A study of remembered context for information access from personal digital archives. In *IIiX 2008 - 2nd International Symposium on Information Interaction in Context*, 2008.
[12] M. Naaman, S. Harada, Q. Wang, H. Garcia-Molina, and A. Paepcke. Context data in geo-referenced digital photo collections. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 196–203, New York, NY, USA, 2004.
[13] K. O'Hara, R. Morris, N. Shadbolt, G. J. Hitch, W. Hall, and N. Beagrie. Memories for life: A review of the science and technology. *Journal of the Royal Society Interface*, 3(8):351–365, June 2006.
[14] N. O'Hare, C. Gurrin, G. J. F. Jones, H. Lee, N. E. O'Connor, and A. F. Smeaton. Using text search for personal photo collections with the mediassist system. In *SAC '07: Proceedings of the 2007 ACM symposium on Applied computing*, pages 880–881, New York, NY, USA, 2007. ACM.
[15] A. J. Sellen, A. Fogg, M. Aitken, S. Hodges, C. Rother, and K. Wood. Do life-logging technologies support memory for the past?: an experimental study using sensecam. In *CHI'07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages81–90, New York, NY, USA, 2007. ACM.
[16] R. S. Wedemann, L. A. V. de Carvalho, and R. Donangelo. A hierarchical memory model for conscious and unconscious mental processes. In *SBRN '06: Proceedings of the Ninth Brazilian Symposium on Neural Networks*, page 14, Washington, DC, USA, 2006. IEEE Computer Society.
[17] Hori T. & Aizawa K. (2003). Context-based video retrieval system for the life-log applications. In *ACM Workshop on Multimedia Information Retrieval*, 31-38.
[18] Jesse E. Purdy, M. r. M., Bennett L. Schwartz, William C Gordon (2001). *Learning and Memory*. California : Wadsworth, Belmont. p. 231-289
[19] Rao, V. and M. Howard, Retrieved context and the discovery of semantic structure, in *Advances in Neural Information Processing Systems*, J.C.P.a.D.K.a.Y.S.a.S. Roweis. 2008, MIT Press: Cambridge, MA. p.1193--1200.
[20] Howard,M.W., & Kahana,M. J. (2002a). A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46, 269-299.
[21] Schwartz, G., Howard, M. W., Jing, B., & Kahana, M. J. (2005). Shadows of the past: Temporal retrieval effects in recognition memory. *Psychological Science*, 16(11), 898-904

# Automatically Adapting the Context of an Intranet Query

Deirdre Lungley
Dept. of Computing and Electronic Systems
University of Essex
Wivenhoe Park, Colchester, CO4 3SQ, UK
*dmlung@essex.ac.uk*

**Abstract**

**In this paper we propose the use of an interactive interface to allow user exploration of the context of an intranet query. The underlying domain model is that of a Formal Concept Analysis (FCA) lattice. Understanding the difficulty of achieving optimum document descriptors, essential for a browsable lattice, we propose harnessing implicit user feedback in learning document/term associations.**

*Keywords: information retrieval, implicit user feedback, query refinement, context enhancement, formal concept analysis, collaborative indexing*

## 1. INTRODUCTION

In an age when companies and institutions are becoming ever more reliant on their intranets, when successful management of their information base is key to efficiency and competitive advantage, the tools they use to access this information are being seen as ever more important. It is reasonable to assume that intranet users are generally very familiar with the traditional search interface, they enter their query term(s) in the text box and scan the resulting document snippets. Many are aware of its fundamental ”bag-of-words” model and have a well-defined search need, which they represent carefully with chosen keywords. The power of modern search technology is such that their need is met in a high proportion of cases. We, in this research, wish to argue the advantages of improving this interface, of going beyond this standard list of results, of providing an interface which illustrates the contexts within which their query term can be found. This enhanced interface is targeted in particular at multi-context and ambiguous queries. Here an interface which guides the user in his choice of terms can reduce the user effort required.

Context enhancement is not new to Information Retrieval (IR). As far back as the early 1970s the use of clustering was promoted as a form of context enhancement (Jardine & van Rijsbergen, 1971). Since then there have been many examples within IR of a query context model being used to enhance the user experience (Sanderson & Croft, 1999, Carpineto & Romano, 2004). In each case researchers have recognised the advantages of offering an enhanced search interface, particularly to less experienced users and users with an information need which is not clearly defined. These enhanced interfaces, illustrating the context of a query, can guide the user in selecting the keywords in use within the collection.

The question then arises as to how accurately the document contexts are reflected in these automatically generated models? These systems attempt to fully automate the context modelling process - a non-trivial task. An alternative would be to use a human defined ontology. We and the above researchers have avoided this option due to its setup and ongoing maintenance costs (Maedche et al., 2003). However, the success of the commercial systems, Aquabrowser[1] and Clusty[2], and evaluation studies of the research systems (Sanderson & Croft, 1999, Carpineto & Romano, 2004) show the potential of query enhancement. Therefore, we set out with this

[1] http://aqua.obeindhoven.nl/
[2] http://clusty.com

research to build on this success, but harness machine learning techniques to achieve further query enhancement - to tune the context model of a user query.

Our research is to be contained within the field of intranet search. We would argue that there is a particular case, both for providing context enhancement and for harnessing user population feedback, in an intranet setting. Since:

- Intranets are relatively controlled environments, consisting of a relatively confined document collection, often with imposed annotation standards and are also far more unlikely to suffer from spamming, making indexing components such as inlinks and metadata more reliable (Fagin et al., 2003). These properties help to ensure the success of the non-trivial task of creating and tuning a query specific context model.
- The user population of an intranet search system is a particular community of users, e.g., the staff and students of a university or the management and employees of a company. The cohesive nature of these communities and their search needs aids the viability of harnessing user population feedback.

A motivating factor in our research has been our analysis of query log data recorded on the University of Essex intranet search engine from the past year. One of the more striking findings of our analysis is the brevity of user search queries. This is in-line with published findings (Jansen et al., 2000). An analysis of the results returned by even such brief queries, shows that the information need of the user, in as far as can be ascertained by the query term, appears to be generally met within the top five results, e.g. graduation, library and timetable. This, however, is not the case when the term is more general, e.g. parking, sport or printing. These more general terms and ambiguous terms like ”union” highlight the potential for context enhancement:

- Printing - Within the university the term printing can have several contexts, e.g. printing centre, printing credit or laboratory printing facilities
- Sport - Again the many contexts of this term could include, sports clubs for children, information on facilities at the Sports Centre or sport science courses. There can be a temporal dimension to a particular context, e.g. in January many parents book summer sports clubs for their children
- Parking - This particular term is often amended to ”parking permit”, a query where the seemingly most relevant document is accessed indirectly through a link in a returned document

The aim of our research is to meet the shortcomings of present search facilities for these types of queries: to provide context information, to adapt this context to temporal requirements and to override inadequate document annotation, by learning a document’s context where necessary. This is to be achieved by learning the optimum association between a document and its terms from past user behaviour.

## 2. RESEARCH QUESTION

We are conducting research designed to answer the following question:

1. Can the context model, the domain specific concept lattice, of an intranet search query be effectively adapted by user implicit feedback?

## 3. PROPOSED RESEARCH METHOD

Many researchers setting out to improve intranet search through context enhancement have followed a common methodology. They have fed the top 100 to 500 documents returned by an underlying search engine into a domain modelling tool (Carpineto & Romano, 2004, Sanderson & Croft, 1999). We aim to follow this common approach, therefore we are adopting the architecture outlined in Figure 1. This illustrates how documents returned by the underlying search engine, in this case Lucene’s Nutch[3], undergo some natural language processing (NLP) in order to extract

[3]http://lucene.apache.org/nutch/

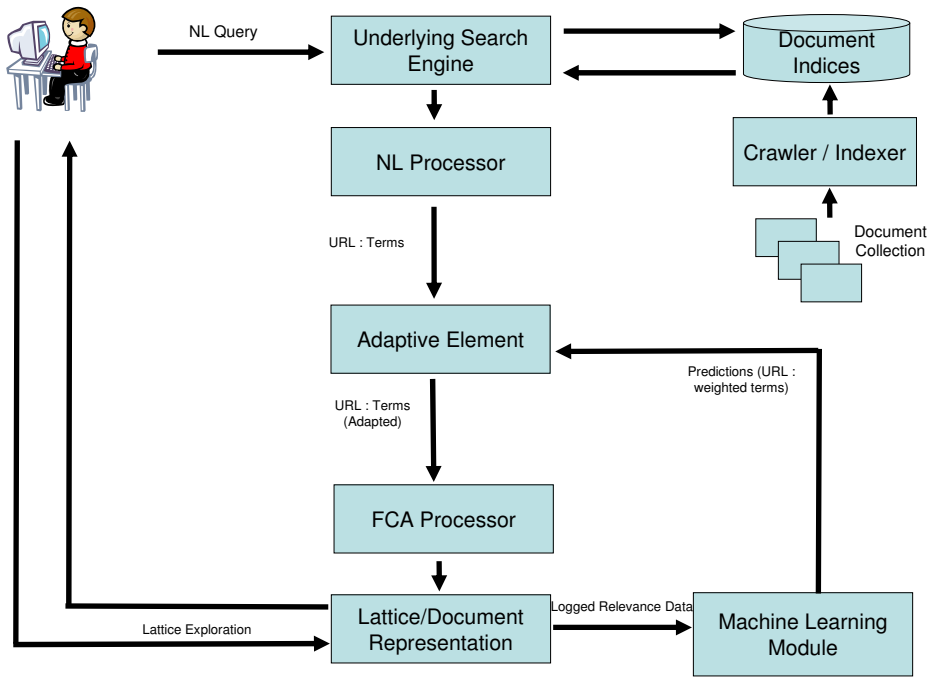



**Figure 1:** System Architecture

initial concept terms. These initial concepts are then merged with the concepts derived from our adaptive element, before being processed by Formal Concept Analysis (FCA) algorithms to produce a browsable concept lattice.

The attraction of the interconnected lattice structure over the more traditional hierarchical tree structure persuaded us to explore FCA as the underlying model of our query context. Toscanaj[4] an open source FCA tool provides us with the algorithms to generate lattice context models from the search engine results. We are not proposing to visualise the full dimensionality of the lattice, but by showing the intent and extent of a node, we are displaying the various interconnections between the derived concepts. The focus of our research is not concerned with FCA. We purely wish to exploit the navigational advantages of the concept lattice. However, a brief description of the terminology of FCA is necessary and is provided in Section 4.

IR researchers interested in exploiting the multi-dimensional structure of the lattice have discovered challenges in using this technique for domain modelling (Cigarrán et al., 2005):

1. Similarly to LSI (Latent Semantic Indexing) FCA is computationally too expensive for large scale processing
2. Unless great care is taken in the selection of index terms, the resulting lattice can be too big and complex for practical browsing
3. Effective visualisation of the lattice is required to harness the potential of this structure

Our research is taking certain steps to surmount these challenges. Firstly we are not setting out to model the entire document collection, but a selection or 100 or so documents returned as results from an underlying search engine. The NL processing we are doing is our answer to the second challenge. This processing must optimise the initial document descriptors, to create a well interconnected lattice, suitable for browsing. At this early stage of our research we are using relatively unsophisticated NLP to extract initial document derived concepts. We are applying a parts-of-speech tagger to each document snippet and using simple noun phrase patterns to locate noun phrases, e.g., noun-preposition-noun. As our research progresses more sophisticated processing might be found beneficial. With regard to lattice visualisation, we are not intending to

[4] http://toscanaj.sourceforge.net/

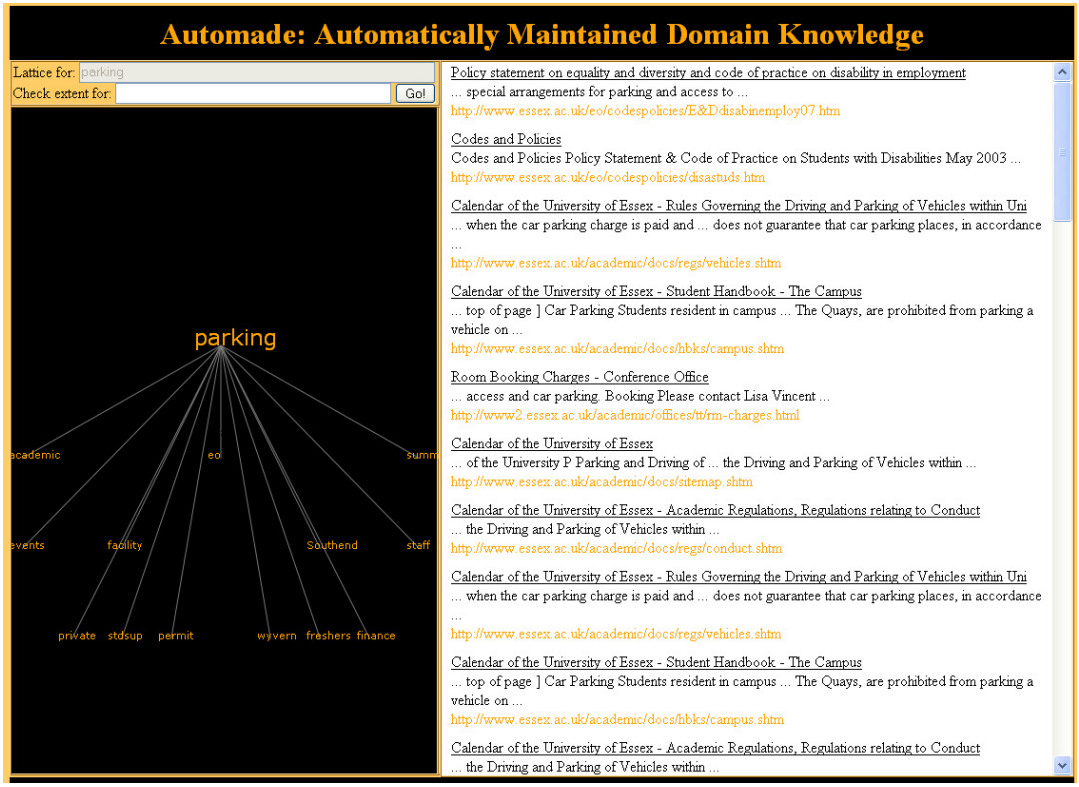



**Figure 2:** Initial extent of unadapted lattice following the query ”parking”

display the complete underlying lattice, but instead to display the highlighted concept, its intent and the concepts within its extent. Figure 2 is an example of our interface.

The core of our research is to automatically improve the context model of a query by harnessing implicit user population feedback (depicted as ”Machine Learning Module” in Figure 1). This will be achieved by learning the document descriptors, the document/term associations. Our initial machine learning processing will utilise Joachims' SVM-Light[5]. Its ranking SVM has been used effectively to re-rank search results (Radlinski & Joachims, 2005). We will explore its potential for tuning our lattice - learning the key documents for a query and the key concepts represented by those documents. Our logged clickthrough data will provide the training data required. Periodic processing of this data will update a model for classification. Figure 2 shows the initial extent of the lattice. The non-adapted lattice appears to give prominence to some terms which appear to bear little relevance to the ”parking” query. Through learning the underlying lattice for this query should become more compact, retaining only relevant documents and relevant terms within those documents.

Our interactive environment allows for continual adaptation. Query terms entered, not derived from the underlying documents, allow new user-driven terms to be added to the context. Logging URL clicks beyond the result documents allows for user-driven documents to be added to the context. Temporal aspects of our context e.g., the autumn teaching timetable is most likely to be the required document for ”timetable” queries at the beginning of October, can be learnt and can subsequently be overridden. It could be argued that learning such temporal features without introducing incorrect associations is a challenging proposal and may benefit from manual intervention. However, our main aim of full automation may mean we have to tolerate a degree of noise. Users, it appears, can tolerate a degree of noise (White & Ruthven, 2006). The selection of a suitable machine learning threshold will be imperative here.

[5]http://svmlight.joachims.org/

| | tch_aut.xls | tch_spr.xls | tch_sum.xls | exam_aut.xls | exam_sum.xls |
|---|---|---|---|---|---|
| timetable | x | x | x | x | x |
| teaching | x | x | x | | |
| exam | | | | x | x |
| autumn | x | | | x | |
| spring | | x | | | |
| summer | | | x | | x |

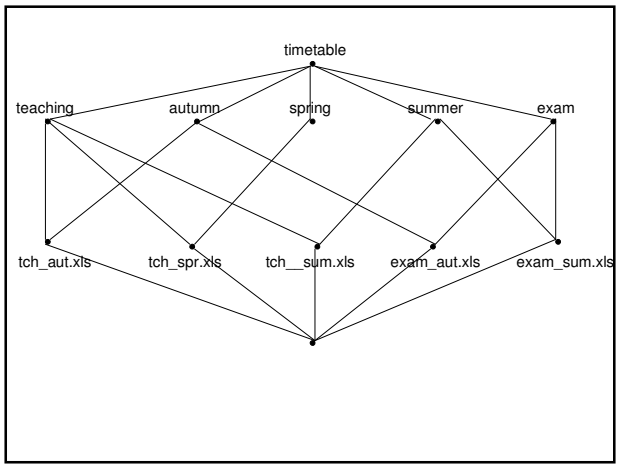



**Figure 3:** Cross-table and related concept lattice

This research aims to evaluate the feasibility of users interactively adapting the context model, the concept lattice, of an intranet query and our final evaluation would therefore involve full interactive IR evaluation. However, our initial explorations of the feasibility of our methodology will involve utilising data from the University of Essex intranet logs. These logs contain the initial query entered and subsequent modifications. We will choose frequently modified multi-context terms e.g., parking, printing and sport, and enter the initial query and subsequent modifications. Our intranet logs do not record the clicked URLs and so our initial investigations will have to make a subjective decision when choosing to click on a URL. The result of these initial investigations will give us some early indications of how effective this combination of Lucene's Nutch search engine, simple NL processing, an FCA domain model and SVM-Light's ranking algorithm are in adapting query contexts.

## 4. FORMAL CONCEPT ANALYSIS (FCA)

The mathematical foundations of FCA have been detailed in Ganter & Wille (1999). The following brief synopsis highlights those aspects which make FCA an attractive option for our domain model.

### 4.1. Contexts and Concepts

In FCA, the extension is the set of objects(entities) belonging to the concept, while the intension is the set of attributes (properties) of these objects. Therefore the basic structure of FCA is the **context** (*G,M,I*), where *I* is a binary relation between the set of **objects** *G* (for the German word Gegenstände) and the set of **attributes** *M* (for the German word Merkmale), $I \subseteq G \times M$. This context can be illustrated by a cross-table as in Figure 3.

We can define *A* as a subset of our object set *G* and *A'* as the set of common attributes of *A*. Similarly, we can define *B* as a subset of our attribute set *M* and *B'* as the set of common objects of *B*. Then the pair *(A,B)* forms a **concept** of the context *(G,M,I)* if *A' = B* and *B' = A*. In other words, in a concept *(A,B)* the set of objects that the members of *B* have in common is *A* and the set of attributes that the members of *A* have in common is *B*. In a concept *(A,B)*, *A*(the set of objects) is called the concept's **extent** and *B*(the set of attributes) is called the concept's **intent**. For example, ({exam_aut.xls,exam_sum.xls},{exam}) is a concept of our context, i.e., exam_aut.xls, exam_sum.xls are the members of the extent; exam is the one member of the intent.

### 4.2. Concept Lattice

We can define a binary relation $\leq$ - "is a subconcept of". If $(A_1, B_1)$ and $(A_2, B_2)$ are concepts of our context, then $(A_1, B_1) \leq (A_2, B_2)$, i.e., $(A_1, B_1)$ is a **subconcept** of $(A_2, B_2)$ if $A_1$ is a subset of $A_2$ and $B_2$ is a subset of $B_1$. $(A_2, B_2)$ is then also called a **superconcept** of the concept $(A_1, B_1)$.

By convention, in the line diagram of a concept lattice, the object names are written below the circle representing the object concept - the object concept being the smallest concept having the object in its extent. The attribute names are written above the circle representing the attribute concept - the largest concept having the attribute in its intent.

## 5. RELATED WORK

Content enhancement e.g. document clustering or classification, has been explored through the decades of IR research in an attempt to organise the results returned by a search engine and to aid the navigation of the user through the information space (Jardine & van Rijsbergen, 1971). It can help the user to clarify the context of his query. The technique of clustering documents into a hierarchy of concept clusters has proved an effective Information Retrieval mechanism as can be seen by the success of the commercial search engine Clusty[6], developed by Vivisimo Ltd. However it is not without its problems. Firstly, it relies on an accurate method of determining the concept for a cluster of documents. Secondly, its hierarchical structure does not inherently allow a user to navigate easily between clusters. IR classification generally involves the use of a term hierarchy and much research is currently being done on their use in ordering documents. Some are investigating the use of large external taxonomies such as the Open Directory Project (ODP)[7]. Although the advantage of communities committing to specific ontologies has its appeal, drawbacks also exist. Ontologies are difficult to maintain (Maedche et al., 2003) and also enforce a specific vocabulary on the user community which has not evolved naturally.

Most IR systems require the user to express their information need in the form of concise query terms and therefore a knowledge of the domain vocabulary is useful (Furnas et al., 1987). Query refinement, another prominent aspect of IR research, can address this issue (Rocchio, 1971).

Formal Concept Analysis (FCA) has been an attempt at combining context enhancement and query refinement: retrieval results organised in a browsable concept lattice, prompt the user in the selection of terms in use within the document collection (Carpineto & Romano, 2004). Over recent years there has been a growing interest among information scientists in FCA and its merits for supporting search have been highlighted by other groups (Cole et al., 2003, Ferré & Ridoux, 2000). Although all these works show the non-trivial nature of using a lattice structure as the basis of a domain model, we are attracted by its navigational capabilities and the prospect of surmounting its difficulties, particularly the choice of index terms, through the use of implicit user feedback.

The use of learning from both implicit and explicit feedback is not new in information retrieval. Relevance feedback has been promoted as an effective method of improving retrieval performance (Rocchio, 1971). Recently, there has been a renewed surge of interest in harnessing the potential of user feedback in an attempt to capture the true information need of the user. Implicit user feedback i.e., user actions such as accessing or printing a document, is seen as an answer to the enormous drawback of explicit feedback, the reluctance of the typical user to rate a document (Jansen et al., 2000). This current interest in using clickthrough data to mine preferences is not without its dissenters. It has been proposed that clickthrough data is not necessarily a good indicator of relevance and one needs to be careful when interpreting this data (Scholer et al., 2008). An answer to this has been to use clicks not as indicators of absolute relevance but of relative relevance, e.g. if a document is clicked on in response to a query, that document is deemed more relevant to that particular query than those listed above it, which have been ignored and the one below it, which has also been ignored (Radlinski & Joachims, 2005). They have had considerable success in using this method to re-rank result documents. Our initial explorations in implicit feedback will also use this concept of relative relevance, not for ranking, but to tune the concept lattice. We are, however, aware of its limitations. Poor content snippets mean that users could repeatedly click on a non-relevant item, so reinforcing invalid relevance and introducing noise. One solution to this is the use of multiple indicators of relevance (Melucci

[6] http://clusty.com
[7] www.dmoz.org

& White, 2007). As our research develops it will be necessary to further explore and employ such discriminating methods.

While some have shown the potential of user population feedback in re-ranking search results (Radlinski & Joachims, 2005, Smyth et al., 2004), others use this technique to improve document retrieval through automatic query expansion (Cui et al., 2002). The language modelling IR research community are also showing considerable interest in feedback. One group derive their query language model by combining the current query terms, related query term histories and related clicked document histories and use it to better the ranking of documents (Shen et al., 2005). Another example of the use of implicit user feedback is where query recommendations are made based on the previous query reformulation behaviour of users (Jones et al., 2006).

Using implicit user feedback to aid context enhancement has also been explored. Web search logs have been mined in an attempt to surmount the classic clustering problems: clusters discovered do not necessarily correspond to the interesting aspects of a topic from the user's perspective and cluster labels generated are not informative enough to allow a user to identify the right cluster (Wang & Zhai, 2007). In answer to these problems they learn these aspects from Web search logs and organise search results accordingly. Search queries and their clicked results have also been used to provide valuable feedback about the relevance of documents to queries (Poblete & Baeza-Yates, 2008). They choose the document's features from the terms of the queries from which it was clicked from and so reduce by over 90% the set of features needed to represent a set of documents. They propose this document representation model for clustering and classification. This use of implicit feedback in aiding the choice of document features is similar to our research, however it aims at reorganising the results rather than adapting a domain model to assist users in interactive search.

The question may be asked as to whether users actually want any refinement/context navigation? Previous task-based evaluations on the University of Essex intranet suggest that users do in fact prefer system-assisted search to a standard search engine (Kruschwitz & Al-Bakour, 2005). It is also interesting to note that mainstream search engines are moving in this direction: Google is making context dependent suggestions, Microsoft Live search is offering related searches for ambiguous queries such as "java" and Yahoo! is offering a "Search Assist Settings" tab.

## 6. ACKNOWLEDGEMENTS

I would like to thank my supervisor Udo Kruschwitz for his help and advice on drafting this paper and two anonymous reviewers for very helpful feedback on an earlier version of this document.

## References

Carpineto, C. & Romano, G. (2004), 'Exploiting the potential of concept lattices for information retrieval with credo', *Journal of Universal Computer Science* **10**(8), 985–1013.

Cigarrán, J., Peñas, A., Gonzalo, J. & Verdejo, F. (2005), Automatic selection of noun phrases as document descriptors in an fca-based information retrieval system, *in* 'Formal Concepts Analysis. Third International Conference, ICFCA 2005', Springer.

Cole, R., Eklund, P. & Stumme, G. (2003), 'Document retrieval for e-mail search and discovery using formal concept analysis', *Applied Artificial Intelligence* **17**(3), 257–280.

Cui, H., Wen, J., Nie, J. & Ma, W. (2002), Probabilistic query expansion using query logs, *in* 'WWW '02: Proceedings of the 11th international conference on World Wide Web', ACM Press, New York, NY, USA, pp. 325–332.

Fagin, R., Kumar, R., McCurley, K., Novak, J., Sivakumar, D., Tomlin, J. & Williamson, D. (2003), Searching the workplace web, *in* 'Proceedings of the Twelfth International World Wide Web Conference', pp. 366–375.

Ferré, S. & Ridoux, O. (2000), A file system based on concept analysis., *in* 'Proceedings of the 1st International Conference on Computational Logic', pp. 1033–1047.

Furnas, G., Landauer, T., Gomez, L. & Dumais, S. (1987), 'The vocabulary problem in human-system communication.', *Comm. of the ACM* **30**(11), 964–971.

Ganter, B. & Wille, R. (1999), *Formal Concept Analysis. Mathematical Foundations.*, Berlin: Springer.

Jansen, B., Spink, A. & Saracevic, T. (2000), 'Real life, real users, and real needs: a study and analysis of user queries on the web.', *Information Processing and Management* **36**(2), 207–227.

Jardine, N. & van Rijsbergen, C. J. (1971), 'The use of hierarchic clustering in information retrieval', *Information Storage and Retrieval* **7**, 217–240.

Jones, R., Rey, B., Madani, O. & Greiner, W. (2006), Generating query substitutions, *in* 'Proceedings of the 15th international conference on World Wide Web', pp. 387–396.

Kruschwitz, U. & Al-Bakour, H. (2005), 'Users want more sophisticated search assistants - results of a task-based evaluation.', *Journal of the American Society for Information Science and Technology (JASIST)* **56**(13), 1377–1393.

Maedche, A., Motik, B., Stojanovic, L., Studer, R. & Volz, R. (2003), An infrastructure for searching, reusing and evolving distrbuted ontologies., *in* 'Proceedings of the Twelfth International World Wide Web Conference', pp. 439–448.

Melucci, M. & White, R. (2007), Utilizing a geometry of context for enhanced implicit feedback, *in* 'Proceedings of the Conference on Information and Knowledge Management', pp. 273–282.

Poblete, B. & Baeza-Yates, R. (2008), Query-sets: Using implicit feedback and query patterns to organize web documents, *in* 'Proceeding of the 17th international conference on World Wide Web', pp. 41–50.

Radlinski, F. & Joachims, T. (2005), Query chains: learning to rank from implicit feedback, *in* 'Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, Chicago, Illinois, USA', pp. 239–248.

Rocchio, J. (1971), *The SMART Retrieval System: Experiments in Automatic Indexing.*, Prentice Hall, Englewood Cliffs, NJ., chapter Relevance Feedback in information retrieval.

Sanderson, M. & Croft, B. (1999), Deriving concept hierarchies from text, *in* 'SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval', ACM Press, New York, NY, USA, pp. 206–213.

Scholer, F., Shokouhi, M., Billerbeck, B. & Turpin, A. (2008), Using clicks as implicit judgements: Expectations versus observations., *in* 'Proceedings of the 30th European Conference on Information Retrieval (ECIR'08)', Glasgow, pp. 52–64.

Shen, X., Tan, B. & Zhai, C. (2005), Context-sensitive information retrieval using implicit feedback, *in* 'Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval', pp. 43–50.

Smyth, B., Balfe, E., Freyne, J., Briggs, P., Coyle, M. & Boydell, O. (2004), 'Exploiting Query Repetition and Regularity in an Adaptive Community-Based Web Search Engine.', *User Modeling and User-Adapted Interaction* **14**(5), 383–423.

Wang, X. & Zhai, C. (2007), Learn from web search logs to organize search results, *in* 'Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval', pp. 87–94.

White, R. W. & Ruthven, I. (2006), 'A study of interface support mechanisms for interactive information retrieval', *Journal of the American Society for Information Science and Technology (JASIST)* **57**(7), 933–948.

# Towards a better understanding of language model information retrieval

M. van der Heijden
Radboud University Nijmegen
*m.vanderheijden@gmail.com*

I.G. Sprinkhuizen-Kuyper
Radboud University Nijmegen
Donders Institute for Brain Cognition and Behavior

Th.P. van der Weide
Radboud University Nijmegen
Institute for Computing and Information Science

**Abstract**

**Language models form a class of successful probabilistic models in information retrieval. However, knowledge of why some methods perform better than others in a particular situation remains limited. In this study we analyze what language model factors influence information retrieval performance. Starting from popular smoothing methods we review what data features have been used. Document length and a measure of document word distribution turned out to be the important factors, in addition to a distinction in estimating the probability of seen and unseen words. We propose a class of parameter-free smoothing methods, of which multiple specific instances are possible. Instead of parameter tuning however, an analysis of data features should be used to decide upon a specific method. Finally, we discuss some initial experiments.**

## 1. INTRODUCTION

Since the end of the last decade language models have caught on as a successful technique for information retrieval. Language models are a specific form of probabilistic models. These models have shown performance similar to vector space models [8], but with more theoretical background compared to the heuristics of vector space [12]. By using explicit models of the query and the documents the representation can be made more precise. This should lead to better performance than a 'bag of words' approach.

In language modeling the most straightforward approach, query likelihood, estimates the relevance of a document by computing the probability of generating the query from the document (e.g. [8]). An alternative and popular approach [4, 10] assumes that the query and documents are each generated from a probability distribution. Similarity is computed by comparing the query and document distributions using relative entropy. Smoothing the document distribution with collection data decreases the influence of data sparseness and query imperfections (see Section 3).

Simple unigram language models assume word independence, although taking preceding words into account improves performance [8]. However, word position is not necessarily a good indication of word dependency. More sophisticated context analysis like non-adjacent word dependency [3] and word relations extracted from external sources [1], has therefore proven interesting. In addition other extensions to language models have been developed, especially probabilistic query expansion techniques [4, 5]. Also document smoothing with a cluster of similar documents has been proposed [6, 9]. These clusters decrease data sparseness because they contain more alternative representations of concepts (e.g. synonyms).

The goal of this study is to understand the mechanics of language modeling and the interactions between various techniques. Although some methods usually perform better than others, for specific situations it is not always clear what method would give the best performance. This results in ad hoc parameter tuning or brute force trial and error of alternative techniques. It is desirable to have better insight in what properties of data and language modeling techniques influence retrieval performance and how these properties interact.

In this study we hope to shed some light on the conditions for optimal performance. In order to do so we will first give an overview in unified notation of language modeling methods from literature. Then in Section 3 we will look specifically at smoothing techniques and the factors that influence parameter settings. New parameter-free smoothing methods will be introduced in Section 4. These are based on the analysis of existing techniques, in an attempt to prevent ad-hoc parameter tuning. We have done some preliminary testing of our work, which can be found in

Sections 5 and 6. From these analyses and experiments we will try to derive future directions for a better theoretical understanding in order to increase language model retrieval performance.

## 2. A LANGUAGE MODELING FRAMEWORK

In order to find relevant documents for a query, model similarity will be used. The query is represented by a probabilistic model $\theta_{qorg}$ and each document by a document model $\theta_{dorg}$. To calculate the similarity between these probabilistic models a similarity measure is needed. A popular measure is relative entropy - also known as Kullback-Leibler divergence. In general, for two (discrete) probability distributions P and Q, relative entropy is written as:

$$D(P,Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

Because this measure is not symmetric there are two possible choices for P and Q. We use the query model as P and the document model as Q. The query and document models can be expanded to overcome data sparseness or to add information not present in the original models. These additions are modeled as a smoothing model $\theta_{qs}$ for the query and a smoothing model $\theta_{ds}$ for a document. Computing the similarity between the models requires interpolating the original and smoothing models, for which we introduce a query interpolation-factor $\lambda_q$ and a document interpolation-factor $\lambda_d$. The similarity between a query and document can now be written as the divergence between the smoothed query and document models:

$$D(\theta_q, \theta_d) = D((1-\lambda_q)\theta_{qorg} + \lambda_q\theta_{qs}, (1-\lambda_d)\theta_{dorg} + \lambda_d\theta_{ds}) \tag{2.1}$$

In order to incorporate the possibility of word dependence in the models we introduce a link model $L$. Simple language models that assume word independence will have $L = \emptyset$. The query and document models are probabilistic models, so the relative entropy can be expanded to probabilities. The model divergence can be written in terms of the probability of a word $w$ given word dependence $L$ and the query model $\theta_q$ or document model $\theta_d$:

$$D(\theta_q, \theta_d) = -\sum_{w \in V} P(w|L,\theta_q)P(L|\theta_q)P(\theta_q) \log \left(P(w|L,\theta_d)P(L|\theta_d)P(\theta_d)\right) \tag{2.2}$$

The query term in the log has been dropped since it is constant for all documents and thus does nog influence ranking. Usually the priors are also disregarded because they depend on external factors (e.g. user, author). These are difficult to take into account, especially in a lab setting.

Assuming word independence is unrealistic (e.g. 'White House'), therefore links between words are considered, represented by our link model $L$. The simple unigram model, also called query-likelihood model, can be recovered from relative entropy when $L = \emptyset$. An estimate of the query model $\theta_q$ using only the original query (the interpolation-factor $\lambda_q$ = 0) then leads to the probability:

$$P(w|\emptyset,\theta_q) = \frac{1}{|q|}\sum_{t \in q} \delta(w,t)$$

where $\delta(w,t)$ is the indicator function which is 1 when $w = t$ and 0 otherwise. Substituting in Equation (2.2) (dropping priors and with $L = \emptyset$) we regain the negative log-likelihood of the simple unigram model [4].

Other N-gram models do take links between words into account and thus have a nonempty $L$. The link between words is static however (e.g. bigram models take only the previous word as link) and as a consequence the probability $P(L|\theta_q)$ equals 1. $P(w|L,\theta_q)$ can then be written as $P(w|w_{-1}\cdots w_{-(N-1)},\theta_q)$. More complex links with non-static probabilities are possible, but they are not further considered here.

Query expansion can be considered a form of query smoothing. Information is added to the query in an attempt to overcome query incompleteness. The factor $\lambda_q$ is a combining weight for the original query and the smoothing model. Whether a word $w$ is a good candidate for expansion can be estimated by calculating the probability of observing that word given the query $q$ and a set of documents $D$ (assumed) relevant to the query [4, 5]:

$$P(w|q) = P(w)\prod_{t \in q}\sum_{d \in D(q)} P(t|d)P(d|w) \tag{2.3}$$

For each query term $t$ the evidence of a relation between $t$ and $w$ is estimated by multiplying the probability of encountering a document $d \in D(q)$ given $w$ and the probability of the query term $t$ given the document $d$. By summing over all documents this probability indicates whether the candidate word fits the query and can be used for query expansion. The query smoothing model can now be constructed by taking the top ranked words, resulting from computing this expansion probability for all words in $D$. However, the expansion model can also be constructed by retrieving synonyms, hypernyms and/or hyponyms from for instance Wordnet.

Another approach to query modeling is the so called relevance model [5]. Instead of using only the query words, a model of all relevant words would perform better. When no training data is available however, the query is still the best estimate of relevance. The difference with the aforementioned query expansion method is that the smoothing model is the probability distribution of $P(w|q)$ (Equation (2.3)) over the whole vocabulary of $D$, instead of a limited set of expansion words.

## 3. SMOOTHING ESTIMATES

One of the factors influencing language model performance is the smoothing strategy used to cope with data sparseness. When estimating the probability of encountering a query term in a document, using normalized frequency will lead to zero probabilities for words not present in the document. This is clearly undesirable, as the absence of one term does not mean the document is completely irrelevant. Smoothing the document estimate with some background distribution can effectively solve the data sparseness problem. The smoothed probability estimate of a word $w$ can be written as an interpolation of the document probability and the background probability:

$$(1 - \lambda)P(w|d) + \lambda P(w|C) \tag{3.1}$$

where $\lambda$ is an interpolation factor ($\lambda_d$ in Equation (2.1)), $d$ is the document and $C$ the collection. For small $\lambda$ this means the estimate is dominated by the count of the word in the document, while for larger $\lambda$ the estimate is closer to the collection probability, so smoothing increases with increasing $\lambda$.

A myriad of methods has been proposed over the years and for every specific situation the optimal technique can be sought. Chen and Goodman [2] compared a number of methods in the context of speech recognition, resulting in a somewhat more formal understanding of various method's merits. These results do not necessarily carry over to language models in information retrieval, because they are used differently. An overview of often used smoothing techniques in IR is given by Zhai and Lafferty [11], based on the work of [2]. The study compares performance of Jelinek-Mercer, Dirichlet and absolute discount smoothing on various corpora and assesses the influence of the smoothing parameters.

The main question remains what factors determine performance of smoothing techniques. The logical place to start would be to look at existing methods and analyze the parameters and properties. Although finding optimal parameters is usually possible, minimizing the number of parameters should improve the model by making it simpler. Following [11] we will analyze the properties of Jelinek-Mercer, Dirichlet and the absolute discounting methods.

The Jelinek-Mercer (jm) method uses a fixed smoothing parameter $\lambda_{jm}$. Obviously, performance is sensitive to the setting of $\lambda_{jm}$, which indirectly depends on corpus characteristics and document characteristics like document length.

So, another class of smoothing methods directly uses document length ($|d|$). Smoothing decreases with document length, which is intuitively correct. Because more data is available longer documents will result in better estimates. So for two documents $d_1, d_2$ the following holds: $|d_1| > |d_2| \rightarrow \lambda(d_1) < \lambda(d_2)$. Dirichlet (dir) is a popular method that fits into this category:

$$\lambda_{dir} = \frac{1}{\frac{1}{\mu}|d| + 1}$$

The parameter $\mu$ is the scaling factor that dictates when a document is considered 'long'. For small $\mu$ the word probability estimate is dominated by the count in the document. For $\mu \gg |d|$ smoothing will increase with $\lambda_{dir}$ going to 1. So, $\mu$ can be seen as a unity for document length and the overall effectiveness depends on the distribution of document lengths in the corpus. See [7] for effects of $\mu$ on the length of the retrieved documents.

To understand absolute discounting (ad) two separate cases should be considered. First, many words will not occur in the document and thus have a probability of zero. All probability estimates of words that do occur in the document are likely overestimated. As a consequence these probabilities should be discounted. Hence the estimate of the probability $P(w|d)$ is:

$$P(w|d) = \begin{cases} \frac{c(w,d)-\delta}{|d|-\delta|V_d|} & \text{if } w \in d \\ 0 & w \notin d \end{cases}$$

The amount of discount is given by $\delta$, in the range $0 \leq \delta \leq 1$. $V_d$ is the vocabulary of the document $d$, and $c(w, d)$ is the count function counting the number of occurrences of $w$ in $d$. The smoothed estimate is constructed by redistributing the probability mass subtracted of seen words to unseen words. Because for every unique word in the document some probability has been discounted, $\lambda_{ad}$ is the ratio between unique words ($|V_d|$) and the length of the document ($|d|$) scaled by $\delta$:

$$\lambda_{ad} = \delta \, \frac{|V_d|}{|\, d \,|}$$

The parameter $\delta$ governs how much probability mass is redistributed to the collection model. Documents that use many different words will be influenced more by the collection whereas documents with a limited vocabulary will be smoothed less. The number of unique words is thus used as a measure for how focused the topic of the document is.

From analyzing these smoothing methods we can try to distill the relevant smoothing factors. The parameter in Jelinek-Mercer smoothing is usually fitted for a specific document collection and set of training queries thus providing only minimal insight in the relevant properties. A comparison of the sensitivity of $\lambda_{jm}$ when using different collections [11], shows that the optimal value may vary considerably with the collection. Dirichlet and absolute discount smoothing incorporate document length, which is a reasonable addition because shorter documents are more likely to be affected by data sparseness. Dirichlet weights the influence of document length with $\mu$, which is often tuned using training data. Although the optimal setting of $\mu$ is also sensitive to the collection used, it is less so than $\lambda_{jm}$ [11]. Absolute discounting introduces information on the word distribution and differentiates the probability estimates for seen and unseen words.

Zhai and Lafferty note that they have not looked at more sophisticated query modeling techniques which may change the results because of the interaction between the query properties and the smoothing performance. Thus it is interesting to establish what the properties of an optimal smoothing method should be when dedicated query modeling is used. Lavrenko and Croft [5] proposed an explicit relevance model to capture the topic of the query, but Lafferty and Zhai [4] have also worked on more specific query modeling. Section 5 contains some initial work in this direction.

Document length, word distribution and the distinction between seen and unseen words are the main factors used in the smoothing methods we looked at. The next section will introduce some alternative smoothing methods, using these features as a basis in combination with some new ideas. The last part of this article will report the experiments and preliminary results.

## 4. NEW SMOOTHING METHODS

A disadvantage of the smothing methods introduced above is that all need parameter tuning. Here we introduce some new smoothing methods which do not need parameter tuning. The main reason to use smoothing methods is to overcome data sparseness. As this is a property of the data, we may be able to use information from the data directly to find the right smoothing parameter. We propose a class of smoothing methods based on word probabilities. However, to prevent introducing a word dependent parameter the information will be generalized over the vocabulary. The class of methods can be described by the function:

$$\lambda(x) = \frac{1}{|V|} \sum_{w \in V} \frac{1}{1 + x(w)} \tag{4.1}$$

with $x$ a function from words to the non-negative numbers. This equation gives a smoothing factor (in the range $(0, 1]$) that depends only on the document and collection (cf. Section 3). The rest of this chapter will give some smoothing strategies that fit within this class.

The ratio between the probability of a word in the document and in the collection indicates the typicality of the word. We will call this ratio $\alpha(w)$:

$$\alpha(w) = P(w|d)/P(w|C) \tag{4.2}$$

$\lambda(\alpha)$ gives us a measure of the centrality of $d$ in $C$. When using $\lambda(\alpha)$ as smoothing method, smoothing will increase when $\alpha(w) = 0$. As documents with a small vocabulary have many words for which $\alpha(w) = 0$, this should alleviate data sparseness.

The Dirichlet method takes document length into account, but is tuned with a free parameter $\mu$. We will attempt to utilize information in the data to replace the parameter. We can use the same concept as in Eq. (4.2) but scale the probability in the document with a document length factor:

$$\beta(w) = P(w|d)/\rho \tag{4.3}$$

where $\rho$ is the relative document length (i.e. scaled to the range [0,1]): $\rho = \frac{\text{rank}(d)}{|C|}$. The longest document has rank 1, the shortest document rank $|C|$ (documents with equal length will get the same rank). When a word $w$ does not occur in the document $\beta(w) = 0$, but in all other cases $\beta$ is weighted with the document length. The smoothing method $\lambda(\beta)$ increases smoothing when data is sparse and takes document length into account. Long documents will thus be smoothed less than short documents, as in Dirichlet smoothing.

Smoothing with $\lambda(\beta)$ uses $P(w|d)$ as a measure of data sparseness but no longer compares this probability with the collection probability. In order to reintroduce this information $\alpha$ and $\beta$ should be combined. This leads to a probability ratio scaled by document length:

$$\gamma(w) = P(w|d)/(\rho P(w|C)) \tag{4.4}$$

In smoothing with $\lambda(\gamma)$ the collection data will be used when $P(w|d)$ is zero, in all other cases scaling with the collection data and document length occurs. Hence, short documents with atypical word probabilities compared to the collection will be smoothed most.

The rationale of introducing the factor $\alpha$ was that the difference between the document and collection probabilities gives information on the amount of smoothing needed. However a consequence of choosing the ratio is that for each word not in the document smoothing increases ($\alpha(w) = 0$ thus the contribution to $\lambda(\alpha)$ for that particular word is 1). Because the collection will have many more words than any single document, this may introduce quite some noise. Instead we could look at a different measure that also indicates the discrepancy between document and collection probabilities:

$$\delta(w) = \frac{1 - |P(w|d) - P(w|C)|}{\rho} \tag{4.5}$$

with $\rho$ again the relative document length. Because smoothing is needed when the absolute difference between the document and collection probabilities is large, we take one minus the difference.

## 5. EXPERIMENTS

We performed some initial experiments on smoothing and query expansion. The goal of these experiments is to get more insight into the factors influencing the retrieval results, in order to predict what we can expect when using the new smoothing methods introduced in Section 4.

The data set comprised AP88-89, WSJ87-92, ZIFF1-2 on TREC-disks 1&2. This standard data set allows us to compare our results with previous studies on smoothing, specifically [11]. The corpus consists of slightly less than 470 thousand documents with a total disk size of approximately 1.5GB. For the language model implementation we used the Lemur toolkit (www.lemurproject.org). The data was preprocessed by stemming with the Porter stemmer. The TREC topics 1 through 150 were used as test queries. Following [11] we used four different kind of queries derived from these topics: short keyword (i.e. title), long keyword (concept field), short verbose (description) and long verbose (title, description and narrative).

A first and rather trivial test is comparing precision/recall (or a related measure, e.g. Mean Average Precision) for smoothed and non-smoothed language model based retrieval. A more interesting test will be gauging the interaction between smoothing and various query expansion techniques. Results obtained using query expansion can be compared to the results without expansion for the mainstream smoothing methods as reported by [11]. This will enable us to test to what extend query expansion is either complementary or comparable to smoothing.

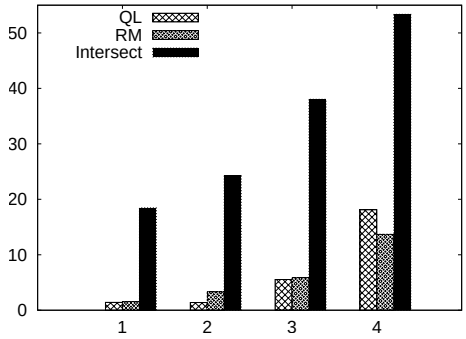



**FIGURE 1:** Histogram showing the number of relevant documents retrieved (y-axis) only by query likelihood (QL), relevance model (RM) and retrieved by both (Intersect) split on document length parts (x-axis, part 1 shortest documents, part 4 longest documents).

Preliminary testing has shown that smoothing methods retrieve different sets of relevant documents, indicating it may be possible to improve retrieval performance by combining methods. The same applies for retrieval with or without query expansion, indicating that it can indeed be seen as a special case of smoothing. Testing will have to prove whether these differences are significant, and we will thus compare the results of various smoothing methods with or without query expansion on the four types of queries.

We hypothesize that we can gain from using a specific smoothing method or a combination of smoothing and query expansion techniques for each type of collection and query. In particular we expect that collections with shorter documents will benefit more from aggressive smoothing and query expansion; longer queries will require less query expansion, but possibly more smoothing to decrease the influence of uninformative words in the query.

In order to find the determining factors for smoothing we divided the collection in parts based on document length. Although document length is a simple measure to divide the collection, it is only a very coarse grained one. Document length does probably not convey useful information on the document content, which is ultimately what we are interested in. Analogous reasoning can be applied to query length.

Instead of looking at division over length it would be more informative to use similarity between the words in queries or documents. We propose a measure to assess word similarity by intersecting the result sets of both words when used as single word queries, summing over word combinations. This results in a score for the cohesion ($\sigma$) within the query or document:

$$\sigma = \frac{2}{N^2 - N} \sum_{i<j}^{N} \text{sim}(w_i, w_j) \quad \text{with} \quad \text{sim}(w_i, w_j) = \frac{|R(w_i) \cap R(w_j)|}{|R(w_i) \cup R(w_j)|} \tag{5.1}$$

and $R(w)$ the result set for a query $w$.

We can use this measure to calculate cohesion within queries and test our smoothing methods against subsets of queries with similar cohesion. The same can be done with documents, but this becomes computationally very expensive because documents are generally longer than queries and the complexity of the cohesion measure is $\mathcal{O}(N^2)$ expensive retrieval operations. Furthermore, comparing large amounts of words will most likely result in summing over coincidental similarities, thus decreasing the information this measure provides.

## 6. RESULTS

A first trivial test is showing that smoothing indeed improves performance. As can be seen in Table 1, retrieval with a simple language model without smoothing (Dirichlet $\mu = 0$) has a Mean Average Precision of around zero. The retrieval has been split over four equally sized parts of the document collection ordered by document length, where part 1 contains the shortest documents and part 4 the longest. Longer documents consistently perform better than shorter documents, which could be a result of longer documents using more words to describe a concept thus providing more possible matches. Furthermore the results indicate that performance is not very sensitive to the setting of the Dirichlet smoothing parameter as performance varies only slightly for parameter settings between 500 and 2500.

| $\mu$ | parts | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 0 | map | .0012 | .0019 | .0044 | .0021 |
| | P5 | .0134 | .0067 | .0121 | .0053 |
| 500 | map | .0243 | .0325 | .0574 | .0938 |
| | P5 | .2604 | .3141 | .3705 | .4200 |
| 1000 | map | .0248 | .0329 | .0590 | .0978 |
| | P5 | .2671 | .3114 | .3919 | .4293 |
| 2500 | map | .0247 | .0329 | .0604 | .1013 |
| | P5 | .2523 | .3047 | .4000 | .4400 |

**TABLE 1:** Mean average precision (map) and precision after 5 documents retrieved (P5) for collection parts (part 1 containing the shortest documents, part 4 the longest documents).

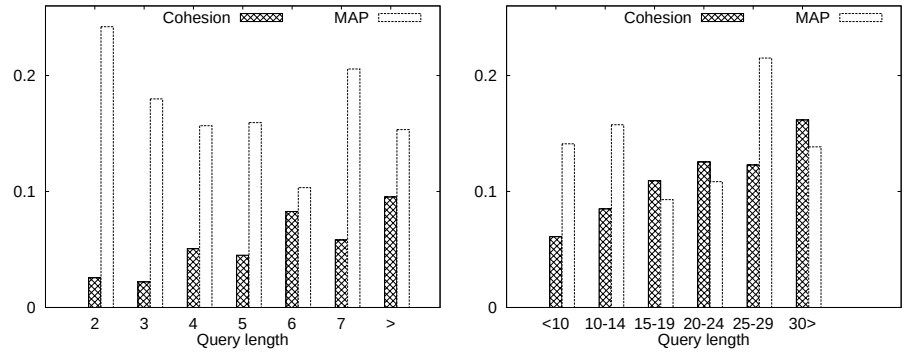



**FIGURE 2:** Cohesion scores and mean average precision (map), for title (left) and description queries (right), binned by query length (x-axis).

We also tested performance for relevance models compared to query likelihood, on different length documents. Figure 1 shows a histogram of relevant documents retrieved by only the query likelihood method (QL), the relevance model (RM) and the overlap between the two. Both methods are able to retrieve documents the other method did not retrieve. An initial attempt to make a rank based combination of results seemed to perform slightly better but this result was not significant, so further testing remains to be done.

Lavrenko and Croft [5] stated that explicitly modeling relevance can address notions of synonymy and polysemy, but apparently it also introduces some noise. Although retrieval with a relevance model on average performs slightly better than without, the differences are small. For short documents performance of QL and RM is very similar, but retrieval without RM produces better results for the longest documents. A possible explanation would be that the RM indeed introduces synonyms that are useful to match short documents with relatively small vocabularies. For longer documents however, it is more likely that the author already used some synonyms enabling better direct matching. The RM would for these longer documents introduce more noise and concept drift decreasing any performance gain.

We were also interested in testing whether smoothing interacted with using a relevance model. Relevance models add words to the query, therefore smoothing may lead to performance loss due to assigning too much probability to less important words. A comparison of a number of smoothing methods against each other and their respective versions with relevance model can be found in Table 2. Lemur-toolkit defaults have been used for the smoothing parameters (i.e. $\lambda_{jm} = .5$, $\mu_{dir} = 1000$, $\delta_{ad} = .7$). Results are furthermore split for query type as described earlier. For the three smoothing methods tested here, mean average precision (map) increases when using a relevance model. Precision after 5 retrieved documents (P5) shows mixed results however. These results indicate that query smoothing by means of a relevance model and document smoothing do not interact. Additionally we can see that differences between smoothing methods are relatively small, but that Dirichlet slightly outperforms the other methods.

In the previous section we introduced a measure for query cohesion/redundancy, in Figure 2 mean cohesion scores are shown binned on query length. Title and description queries tend to have more cohesion when queries are longer. Longer queries with redundancy should show performance similar to shorter queries, how much performance is lost with increasing redundancy still needs further analysis. This could shed some light on whether the lower retrieval scores of description queries (see Table 2) are caused by a length effect (due to redundancy), or a

| Query type | | Title | | Description | | Long keyword | | Verbose | |
|---|---|---|---|---|---|---|---|---|---|
| Method | | QL | RM | QL | RM | QL | RM | QL | RM |
| JM | map | .1571 | .2082 | .1195 | .1598 | .1991 | .2231 | .2071 | .2168 |
| | P5 | .3693 | .4360 | .3280 | .3573 | .4765 | .4591 | .4773 | .4707 |
| Dirichlet | map | .1839 | .2078 | .1365 | .1678 | .2127 | .2317 | .2158 | .2325 |
| | P5 | .4360 | .4480 | .3947 | .3800 | .5221 | .5168 | .5160 | .5053 |
| AbsDiscount | map | .1677 | .2046 | .1164 | .1408 | .1986 | .2236 | .1985 | .2019 |
| | P5 | .4187 | .4640 | .3547 | .3653 | .4926 | .5034 | .4800 | .5107 |

**TABLE 2:** Comparison of performance for Query likelihood (QL) and Relevance model (RM) retrieval with the smoothing methods as described in Section 3.

consequence of more noise in the query due to sentence filler words that are not present in keyword queries.

## 7. CONCLUSION

In this study we have analyzed what factors influence language model information retrieval performance. Starting from popular smoothing methods we established what data features were used. Document length and a measure of document word distribution turned out to be the important factors, in addition to a distinction in estimating the probability of seen and unseen words. We used this information as a basis to develop parameter-free smoothing methods. These proposals should enable us to further analyse how performance is influenced by smoothing.

Replicating [11] we show that Dirichlet smoothing is not very sensitive to the setting of its parameter $\mu$. We can also conclude that longer documents can be ranked better than short documents, most likely because they contain more information. We have also looked at the possibility of interaction between using a relevance model [5] for query expansion and smoothing methods. No direct interaction was found, but we did find that retrieval without a relevance model resulted in different relevant documents than retrieval with a relevance model. This needs further analysis. A combination of the results with and without the relevance model seemed to improve performance, but the improvement was not significant. As such we cannot draw any conclusions from our initial experiments. Future work will consist of experiments that further analyze the characteristics of both the data and the retrieval methods used, further improving the overview of the interaction between language modeling methods and retrieval data.

## REFERENCES


[1] G. Cao, J.-Y. Nie, and J. Bai. Integrating word relationships into language models. In *Proceedings of the ACM SIGIR Conference '05*, pages 298–305, 2005.

[2] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Harvard University, 1998.

[3] J. Gao, J.-Y. Nie, G. Wu, and G. Cao. Dependence language model for information retrieval. In *Proceedings of the ACM SIGIR Conference '04*, pages 170–177, 2004.

[4] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the ACM SIGIR Conference '01*, pages 111–119, 2001.

[5] V. Lavrenko and W. B. Croft. Relevance based language models. In *Proceedings of the ACM SIGIR Conference '01*, pages 120–127, 2001.

[6] X. Liu and W. B. Croft. Cluster-based retrieval using language models. In *Proceedings of the ACM SIGIR Conference '04*, pages 186–193, 2004.

[7] D. E. Losada and L. Azzorpardi. An analysis on document length retrieval trends in language modeling smoothing. *Information Retrieval*, 11:109–138, 2008.

[8] D. R. H. Miller, T. Leek, and R. M. Schwartz. A hidden Markov model information retrieval system. In *Proceedings of the ACM SIGIR Conference '99*, pages 214–221, 1999.

[9] T. Tao, X. Wang, Q. Mei, and C. Zhai. Language model information retrieval with document expansion. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 407–414, 2006.

[10] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, pages 403–410, 2001.

[11] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2), 2004.

[12] C. Zhai and J. D. Lafferty. A risk minimization framework for information retrieval. *Inf. Proces. Man.*, 42:31–55, 2006.

# An Investigation into Query Throughput and Load Balance Using Grid IR

Ahmad Abusukhon[1], Michael P.Oakes[1]
[1]School of Computing and Technology
University of Sunderland,
Chester Road, Sunderland
SR1 3SD, UK
Tel.: +44(0)1915152222
*ahmad.abusukhon@sunderland.ac.uk, michael.oakes@sunderland.ac.uk*

**Grid information retrieval (GIR) means using a grid system for retrieving relevant documents that satisfy the user need from within a large-scale data collection. A data collection can be text, audio, video, etc. The grid provides powerful computation while information retrieval provides techniques for retrieving useful information. In previous work, we built our baseline system and we compared three types of IR systems, namely document-based, term-based and hybrid partitioning with respect to average query response time. In addition, we proposed new methods for improving the load balance and query throughput for term-based and hybrid partitioning. We carried out a set of real experiments using a broker and six nodes. In this paper, we propose to repeat our previous experiments using the grid as a new IR environment.**

*Load balance, query-throughput, query-routing, partitioning methods, query response time, Grid IR.*

## 1. INTRODUCTION

The number of documents available on-line has grown rapidly since the birth of the World Wide Web. Baeza-Yates and Ribeiro-Neto (1999, p.229) mentioned that the web size is growing very fast, nearly doubling in size every six months. Currently, there are more than 20 billion indexed web pages, (Baeza-Yates et al., 2007). Besides the huge number of documents, we have a huge number of queries (billions of queries) submitted by users (Badue et al., 2001). Thus, as the number of queries submitted by the users increases, it becomes necessary to increase the system throughput (the number of queries executed per second), and the IR system must provide a high query processing rate.

T.Meadow et al. (2007) defined Information retrieval as “finding some desired information in a store of information or a database”. They also defined information retrieval as a communication process which includes finding documents, graphic images or sound recordings that are relevant to the user need. Using the information retrieval techniques users can communicate with a librarian, museum or fingerprint identification specialist. Thus, the main aim of an information retrieval (IR) system is to provide users with efficient and easy access to the information relevant to their needs. This task becomes difficult as the size of the data collection becomes large (multi Gbytes).

IR systems are implemented either on a single machine (i.e. sequential IR systems) or on parallel machines. In sequential IR systems, only one machine is responsible for constructing an index and answering the user query. MacFarlane et al. (2005) stated that “indexing large multi-gigabyte text databases can take many hours or even days to complete”. Zobel and Moffat (2006) mentioned that the main problem that made index construction challenging was that the entire index of the whole collection could not fit in main memory. In addition, answering queries using a single machine resulted in low query response time.

To solve the above problems, traditional IR systems must change their architectures and algorithms to parallel and distributed architectures (Baeza-Yates & Ribeiro-Neto, 1999). D.Manning et al. (2008, p.74), mentioned that for large collections (for example the Web) we cannot construct the index on a single machine, so we should use large clusters of computers in order to carry out this task.

In distributed IR systems, a big problem is divided into a small task (i.e. the overall computation is decomposed into a number of small tasks) and then each node is assigned a certain task. The decomposition and assignment of tasks to different nodes in the system is called partitioning. The query response time is greatly influenced by the way in which a large scale data collection is partitioned across multiple nodes. Jeong and Omiecinski (1995) proposed two different methods for partitioning an inverted file among multiple disks, namely document id and term id. One of the major problems that may arise here is the load balance. The problem of load balancing is to develop a partitioning and mapping algorithm in order to balance the workload across nodes and reduce the communication time between the nodes (processors).

In the previous work, we built our baseline system by comparing between three types of IR systems, namely document-based, term-based and hybrid partitioning with respect to the average query response time. In addition, we showed how the load balance and the query throughput of term-based and hybrid partitioning could be improved by using new methods for query routing and collection partitioning (Abusukhon et al. 2008a, b, c), see Sec. 3.

In future, we propose to repeat our previous experiments using the University of Sunderland Cluster Computer (UoSCC) or simply the Grid. The University of Sunderland launched the Grid on 24th of April 2008. The grid consists of central node (Xeon Intel Core Duo 64bit) that communicates with a cluster of nodes called the head nodes. The head nodes are connected to compute nodes (based upon Dell Server type 2950). The total number of nodes in the system is 42 computing nodes. The number of CPU processors per node is two, while the RAM of each node is 8Gbytes. The network bandwidth is 1Gbps, the data storage is a SATA drive, 250Gbyte per node, and thus the total distributed storage is 40 * 250Gb = 10Tb. In addition, the grid uses dual operating systems namely Linux and Windows. The Grid (UoSCC) has been designed in order to be a general purpose machine, and may be applied in different areas. For example, it may be used for solving scientific and engineering problems such as network planning, web based search engines, media applications such as rendering and data transcoding, multimedia and high bandwidth video streaming, general purpose simulation parameter exploration and numerical optimisation, and computational fluid dynamics (CFD).

GIR (Grid Information Retrieval) is similar to distributed computing. It consists of a number of nodes that are connected by a network. In addition, tasks (or subtasks) are assigned to different nodes, and carried out in parallel. However, the difference between GIR and distributed systems is that GIR provides more finely grained implementation for task assignment and coordination among the grid elements. In fact, GIR is composed of three components that can exist on any machine on the computational grid, namely collection managers (CMs), indexers, and query processors (QPs). The collection manager is responsible for accessing, storing, transforming and delivering data items from the collection, and CMs may have privileged access to data. Indexers make up the heart of the traditional IR system; they build the index of a given data collection in order to provide efficient access to that collection. The query processors are responsible for interacting with the indexers, over time, in order to collect query answers, merge them, and send them to their users (Dovey & Gamiel, n.d.; Gamiel et al., 2004).

The performance of IR systems is measured by retrieval efficiency and retrieval effectiveness. Measures such as recall and precision are commonly used for measuring retrieval effectiveness (R.Korfhage 1997, p.194). Suppose that we have a user query Q and its set (R) of relevant documents. Let the number of documents in the set (R) equal r. Assume that the query answer (after applying a given retrieval strategy) to query Q is the set of documents (A). Let the number of documents in this set equal a. In addition, let Ri be the number of documents in the intersection of the sets R and A, then the recall and precision measures are defined as follows. Recall is “the fraction of the relevant documents (the set R), which has been retrieved” i.e., recall = Ri/r. Precision is “the fraction of the retrieved documents (the set A), which is relevant” i.e., precision = Ri/a (Baeza-Yates & Ribeiro-Neto, 1999 p.75). Retrieval efficiency is defined as “the measure of the time taken by an IR system to do a computation on the database, although this usually means search it” MacFarlane (2000). The gain in retrieval efficiency using distributed IR or Grid computation against sequential IR is measured by the speed up and efficiency of the system as well as the system throughput. For example, how many queries are carried out per second? How many Mbytes are indexed per hour?

## 2. RELATED WORK

In IR there are three types of collection partitioning, namely document-based, term-based and hybrid partitioning. In document-based partitioning, documents are distributed across nodes in a circular round robin fashion such that each node gets a sub-collection of the whole data collection. Each node builds its local index for the sub-collection it received and becomes responsible for answering queries about its sub-collection. In document-based partitioning, queries are carried out sequentially. In other words, it is difficult to direct queries to their relevant nodes because all nodes are responsible for answering a certain query term (A.Ribeiro-Neto & A.Barbosa, 1998; Badue et al., 2001;Cambazoglu et al., 2006). In order to achieve better load balance when using document-based partitioning, Ma et al. (2002) proposed three different approaches for inverted list partitioning, namely consecutive partitioning scheme, interleaving scheme, and differential partitioning. In consecutive partitioning, each workstation holds a set of consecutive documents identifiers (ID’s). In interleaved partitioning, the mapping between the document number and the workstation was carried out using the equation L(d) = d Mod M, where d is the document identifier (ID), and M is the number of workstations. In differential partitioning, a document weight was assigned to each document based on the probability that a certain term appears in the query, then each node was assigned a set of documents with balanced-weight using the following equation. Balanced-weight= total-document-weight / M, where M is the number of workstations.

In term-based partitioning, a global index is built where the terms and their complete inverted lists are stored in the broker and then the broker distributes the terms associated with their inverted lists across nodes such that each node gets a sub-set of terms, and thus each node is responsible for answering queries about the terms it stores. In term-based partitioning queries are directed to their relevant nodes and the IR system can carry queries concurrently, where one or multiple queries can be submitted to the system at once (Cambazoglu et al., 2006).

In order to improve the load balance, Xi et al. (2002) proposed hybrid partitioning. They divided the inverted lists into a number of chunks and then distributed them across nodes randomly. The broker distributed the query term across all nodes (i.e. query terms were not directed to their relevant nodes) sequentially (i.e. one query at a time), thus all nodes were responsible for answering a certain query term. They increased the query throughput of their system by increasing the multi-programming level in order to carry out more than one query term at a time. They concluded that hybrid partitioning outperformed document-based partitioning especially when the chunk size was small.

Improving the load balance of term-based partitioning was the focus of other researchers. Current approaches for improving the load balance for term-based partitioning can involve concurrent queries where more than one query is processed at the same time (Cambazoglu et al., 2006). This is done in order to reduce the number of idle nodes in the system, and thus achieve better load balance and increase the system query throughput as well. Alternatively, using partial replication and caching, Moffat et al. (2006) aimed to balance the load for pipelined term-distributed parallel architecture, and they proposed different techniques in order to reduce the query costs. They showed that load imbalance could be addressed by techniques based on predictive inverted list assignments to nodes and selective list replication. In a pipelined system, query evaluation is executed as follows: suppose that we have N nodes, and that we have a query consisting of three terms (t1, t2, and t3) that reside on three nodes: n1, n2, and n3. The query evaluation begins at n1, which retrieves the inverted list of term t1, sends it to node n2, which retrieves the inverted list of t2 and sends it with the inverted list of t1 to n3 and so on. The disadvantage of this system was the load imbalance, which was caused by a number of terms with long inverted lists. In order to solve this problem they proposed replicating the inverted lists of those terms over nodes.

Marin and Costa (2007) investigated improving the load balance for term-based and document-based partitioning by balancing query ranking and query fetching. They mentioned that on current cluster technology the most expensive steps (most dominant factors) in query evaluation are fetching lists from disk and query ranking. Thus, in their proposed system they detached ranking from fetching. List fetch was balanced by caching the most frequent terms in queries. A third technique is to balance query ranking and query fetching (Marin and Costa, 2007).

In fact, searching inverted lists is heavily disk dependent. In other words, most of the searching time is spent on retrieving the inverted lists from the disk. Dewitt & Gray (1990) mentioned that query response time decreased when more processors and disks were used (there is a point beyond which increasing the number of processors and disks increases the query response time), this is because the index size searched at each node becomes smaller and thus the response time decreases. Badue et al. (2001) used shared-nothing architecture in order to compare term-based and document-based partitioning. They concluded that term-based partitioning was better than document-based partitioning because term-based allowed the parallelization of disk seeking. MacFarlane (2000) compared document-based and term-based partitioning and they concluded within their system that document-based is the preferable method for search. Frieder & Tova Siegelmann (1991) concluded that the performance of multiprocessor information retrieval systems depends on both the underlying parallel technology and the organization of the data to be retrieved. They stated that “poor data allocations result in minimal performance gains on a parallel engine as compared to a uniprocessor system”. In other words, different data partitioning (allocation) algorithms lead to different retrieval efficiency (i.e. different query response times).Thus it is very important to decide where the inverted lists must reside.

Partitioning the tasks across nodes is not an easy task. For example, partitioning the index (e.g. term-based index) by sending an equal number of terms across multiple nodes may not always result in equal amount of work (Frieder et al.,2000; Moffat et al., 2006). Although different algorithms have been implemented in order to improve the load balance of term-based and document-based partitioning, this problem is still an open question in distributed IR systems (Baeza-Yates et al., 2007).

Our research looks at the problem of efficiently retrieving (i.e. reducing the query response time) for documents that are relevant to the user needs in a large-scale static data collection (multiple Gigabytes). In our research, we focus on improving the load balance and the query throughput as factors of the query response time.

## 3. WORK COMPLETED TO DATE

### 3.1 Baseline system

Abusukhon et al. (2008a), compared three types of IR systems namely, document-based, term-based and hybrid partitioning with respect to the average query response time. In their experiment they used six nodes (256 RAM) connected to a broker via 10/100 Ethernet switch. They used the data collection WT10G, and 50 queries (451-500) from TREC-9. In document-based and hybrid partitioning, query terms were broadcasted over all nodes. In term-based partitioning, query terms were directed to their relevant nodes. Unlike Xi et al. (2002), they did not find that hybrid partitioning was any better than document-based partitioning in terms of the average query response time although they chose the chunk size to be as small as possible. They justified their conclusion with respect to communication and merging time. However, as in the previous work, they found that document-based partitioning and hybrid partitioning outperformed term-based partitioning in terms of average query response time. In hybrid

partitioning and document-based partitioning a certain query term may appear on more than one node, thus queries were broadcasted over all nodes because the broker does not have any pre-knowledge about where each term and its associated list are located.

### 3.2 An investigation into improving query throughput and load balance of hybrid partitioning

Abusukhon et al. (2008b), mentioned that hybrid partitioning suffered from low query throughput because it was difficult to direct queries to their relevant nodes. In order to solve this problem they proposed to divide the nodes into clusters such that each cluster of nodes is responsible for answering a query term (note that in Xi et al.'s method, all nodes were responsible for answering a certain query term). They built the term-based index and then they divided each inverted list into (m) chunks, where m equals the number of nodes in each cluster. They distributed the chunks associated with their terms across clusters such that all terms started with a certain set of letters resided on a certain cluster. In addition, they distributed the chunks of a certain set of terms (for example, all terms starting with the letters A to D) across the nodes of a specific cluster such that each node in the cluster stored part of the complete inverted list of a certain term. They implemented the hybrid IR system and then they ran 50 queries and measured the load balance and query throughput, and then they implemented their IR system in which queries were directed to their relevant clusters and they measured the query throughput and the load balance.

Their results showed that their proposed method improved the query throughput of Xi et al.'s by 60%. In addition, they investigated carrying out multiple queries (μ queries) at once. They defined the hybrid queries which resulted from combining μ queries into one query (Q) then removing all duplicated terms from (Q) and then split Q into N streams, where each stream contains all terms (from all queries) starting with a certain set of letters (for example, A to D), and then they directed each stream to its relevant cluster (node). They studied how the query throughput and the load balance were affected by μ. They found that increasing the μ value led to better load balance and query throughput and thus the average query response time was reduced. They compared term-based partitioning and hybrid partitioning when μ = 50 queries (carried out at once) and they found that within their system the hybrid partitioning performed better than term-based partitioning. This was the first work which compared term-based and hybrid partitioning when queries were directed to their relevant nodes.

### 3.3 An investigation into improving the load balance of term-based partitioning

The main disadvantage of term-based partitioning was the load imbalance. Directing queries to their relevant nodes results in load imbalance and thus some nodes (those which store the most frequent terms) are heavily loaded while other nodes may be idle or lightly loaded (Badue et al., 2001). Recent research (Baeza et al., 2007), showed that it is still unclear on the circumstances under which each of the two partitioning algorithms (term-based and document-based) is suitable. They also stated that another open problem is how to find an efficient way of data partitioning for both term-based and document-based partitioning that achieves load balance among the different servers.

In order to solve the above problem, Abusukhon et al. (2008c) proposed two new techniques for query routing and index partitioning, namely term-total-frequency and term-length partitioning. In the first technique, they proposed to distribute the term-based index equally across nodes with respect to the total term frequency calculated from the inverted lists. For example, suppose that term (x) appears in the following inverted list: 3:1, 6:3, 9:2, 20:4, where each pair represents the document identifier and the term frequency in that document, then the total term frequency of term x equals 10. They passed through the term-based index and they calculated the total frequency (F) for each term, and then they stored (F) and the term associated with (F) into a lookup file. The lookup file was sorted with respect to the total term frequency and then the term-based index was distributed among nodes with respect to the lookup file.

The aim of the above experiment was to distribute the load (the summation of the total frequency of all terms) equally across nodes. The motivation for their work was the Zipf's law, which is used to capture the distribution of the frequencies of the words in a given text. This law stated that "the frequency of the i-th most frequent word is $1/i^r$ times that the most frequent word", (r being between 1.5 and 2.0), thus the frequency of any word is inversely proportional to its rank (i.e. i-th position) in the frequency table. Their hypothesis was that if the most frequent terms in the term-based index were distributed equally across nodes then the load balance might be improved. In the second technique, they proposed to distribute the term-based index equally among nodes with respect to the term length (in characters).

The motivation for term length partitioning comes from the observation that queries in Excite-97 have a very skewed distribution of term lengths with some predominant lengths. Their hypothesis was that if the predominant-length terms resided on one or two nodes in the system then most of the user query terms will be answered by one or two nodes while other nodes are lightly loaded or idle. Thus, the terms of the term-based index were distributed equally across nodes with respect to the term length in order to improve the load balance. In addition, they studied the load balance of term-based partitioning when terms were distributed among nodes in a circular round robin fashion, as well as, when terms were distributed equally among nodes with respect to the inverted list length (in bytes). They ran 10,000 queries from Excite-97, and they measured the load balance of the above four techniques.

Their results showed that within their system, the term-length partitioning performed better than other techniques in terms of load balance.

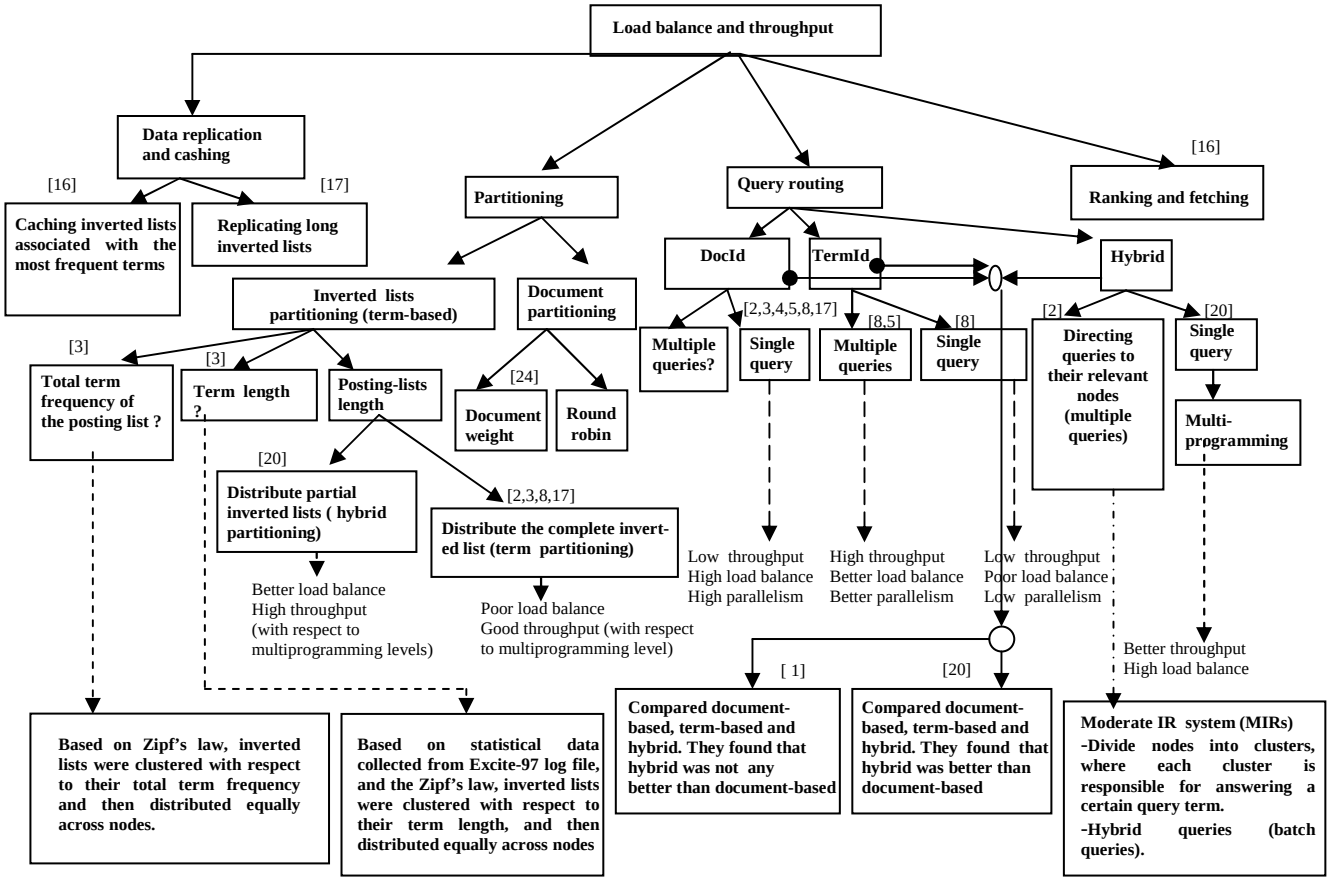



**FIGURE 1:** Analysis of the previous work

## 4. CONCLUSION

Fig. 1 shows a summary of work by previous authors on improving the load balance and the query throughput as important factors of the query response time, to illustrate where our own work (labelled [1], [2] and [3] as in the reference list below) has attempted to fill in the gaps in the literature. Overall, we have addressed the problems of how to improve the query throughput and load balance for hybrid partitioning and how to improve the load balance for term-based partitioning. In future, we will repeat the experiments we have already carried out using 7 nodes, this time using the Grid System at Sunderland. The Grid provides powerful computation because it consists of large number of computational nodes (42) that work in parallel in order to solve large-scale problems.

## 5. ACKNOWLEDGMENT

Our thanks to Al-Zaytoonah University of Jordan and The University of Sunderland for their sponsorship.

REFERENCES.

[1] Abusukhon, A., Oakes, M. Talib, M. and Abdalla, A. (2008a) Comparison Between Document-based, Term-based and Hybrid Partitioning. In Snasel,V. et al. (Eds.), *Proceedings of the First IEEE International Conference on the Application of Digital Information and Web Technologies.* Ostrava, Czech Republic, 4-6 August, pp. 90-95. IEEE.

[2] Abusukhon, A., Talib, M. and Oakes, M. (2008b) Improving the Load Balance for Hybrid Partitioning Scheme by Directing Hybrid Queries. In Burkhart, H. (Eds.), *Proceedings of the IASTED International Conference on Parallel and Distributed Computing and Networks as part of the 26th IASTED International Multi-Conference on APPLIED INFORMATICS.* Innsbruck, Austria, 12-14 February, pp. 238-244. ACTA press, USA.

[3] Abusukhon, A., Talib, M., and Oakes, M. (2008c) An Investigation into Improving the Load Balance for Term-Based Partitioning. In Kaschek, R. et al. (Eds.), *Proceedings of UNISCON 2008, the 2nd International United Information Systems Conference.* Klagenfurt, Austria, 22-25 April. LNBIP 5, pp. 380-392. Springer-Verlag, BerlinHeidelberg-Germany.

[4] Ribeiro-Neto, B., and Barbosa, R. (1998) Query performance for tightly coupled distributed digital libraries. *Proceedings of the third ACM Conference on Digital Libraries*, pp. 182-190. ACM, New York, USA. Available at : http://delivery.acm.org/10.1145/280000/276695/p182-ribeiro-neto.pdf?key1=276695&key2=2932507021&coll=GUIDE&dl=GUIDE&CFID=22405030&CFTOKEN=47383206 [accessed on 01-April-2008].

[5] Badue, C., Baeza-Yates, R., Ribeiro-Neto, B., and Ziviani, N. (2001) Distributed Query Processing Using Partitioned Inverted Files. *Proceeding of the 9th string processing and information retrieval (SPIRE) symposium.* pp. 10-20. IEEE CS press.

[6] Baeza-Yates, R., Castillo, C., Junqueira, F., Plachouras, V., and Silvestri, F. (2007) Challenges on Distributed Web retrieval. *Proceeding of IEEE 23rd International Conference on Data Engineering (ICDE)*, Istanbul, Turkey, 15-20 April, pp. 6-20. Available at: http://www.dcc.uchile.cl/~ccastill/papers/baeza_2007_challengesz_distributed_information_retrieval.pdf. [accessed on 15-February-08].

[7] Baeza-Yates, R. and Ribeiro-Neto, B. (1999) *Modern Information Retrieval*. Addison-Wesley, New York: ACM Press.

[8] Cambazoglu, B., Catal, A., and Aykanat, C. (2006) Effect of Inverted Index Partitioning Schemes on Performance of Query Processing in Parallel Text Retrieval Systems. In Levi, A. et al. (Eds.) *ISCIS 2006*. LNCS, pp. 717-725. Springer, Heidelberg.

[9] Manning, C., Raghavan, P., and Schütze, H. (2008) *An Introduction to Information Retrieval*. [e-book] Cambridge England: Cambridge University Press. Available at: http://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf [accessed on 24-February-08].

[10] Frieder, O., Grossman, D., Chowdhury, A., and Frieder, G. (2000) Efficiency Considerations for Scalable Information Retrieval Servers. *Journal of Digital information*, 1(5).

[11] Gamiel, K., Newby, G., and Nassar, N. (2004) *Grid Information Retrieval Requirements*. Copyright © Global Grid Forum 2004. Available at : http://www.ogf.org/documents/GFD.27.pdf [accessed on 21-June-2008].

[12] Gulli, A., and Signorini, A. (2005) The Indexable Web is More than 11.5 billion pages. *Special interest tracks and posters of the 14th international conference on World Wide Web,* Japan, pp. 902-903. ACM, New York, USA. Available at: at : http://portal.acm.org/citation.cfm?id=1062789. [accessed on 31-March-2008].

[13] Dovey, M., and Gamiel, K. (n.d.) Grid IR- GRID Information Retrieval. [Internet]. Available at: http://www.w3c.rl.ac.uk/Euroweb/poster/112/gridir.html. [accessed on 06-June-08].

[14] MacFarlane, A. (2000) *Distributed inverted files and performance: A study of parallelism and data distribution methods in IR*. Ph.D. City University, London. Available at: http://www.soi.city.ac.uk/~andym/personal.html. [accessed on 14-May-08].

[15] MacFarlane, A., McCann, .J.A., and Robertson, .S.E. (2005) Parallel methods for the generation of partitioned inverted files. *Aslib Proceedings*, 57(5), 434-459.

[16] Marin, M., and Gil Costa, V. (2007) High-Performance Distributed Inverted Files. *Proceedings of the sixteenth ACM conference on CIKM'07*. Lisboa, Portugal, pp. 935-938. ACM, New York, USA.

[17] Moffat, A., Webber, W., and Zobel, J. (2006) Load Balancing for Term-Distributed Parallel Retrieval. *Proceeding of the 29th annual international ACM SIGIR conference on Research and development in information*, pp. 348-355. ACM, New York, USA.

[18] R. Korfhage, R. (1997) *Information Storage and Retrieval*. USA : John Wiley & Sons, Inc.

[19] T.Meadow, C., R.Boyce, B., H.Kraft, D., and Barry,C. (2007) *Text Information Retrieval System*. 3rd ed. London UK: Elsevier.

[20] Xi, W., Somil, O., Luo, M., and Fox, E. (2002) Hybrid partition inverted files for large-scale digital libraries. *Proceedings of Digital Library: IT Opportunities and Challenges in the New Millennium*. Beijing, China. Beijing Library Press.

[21] Zobel, J., and Moffat, A. (2006) Inverted Files for Text Search Engines. *ACM Computing Surveys (CSUR)*, 38(2).

[22] Zobel, J., Moffat, A., and Ramamohanarao, K. (1998) Inverted Files Versus Signature Files for Text Indexing. *ACM Transactions on Database systems*, 23(4), 453–490

[23] Jeong, B., and Omiecinski, E. (1995) Inverted File Partitioning Schemes in Multiple Disk Systems. *IEEE, Transactions on Parallel and Distributed Systems*, 6(2), 142–153.

[24] Ma, Y.-C., Chen, T.-F., and Chung, C.P. (2002) Posting file partitioning and parallel information retrieval. *Journal of systems and software*, 63(2), 113-127, Elsevier.

[25] Dewitt, D.J. and Gray, J. (1990) Parallel database systems: The future of database processing or a passing fad. *SIGMOD RECORD*, 19(4), 104-112.

[26] Frieder, O., and Tova Siegelmann, H. (1991) On the allocation of documents in multiprocessor information retrieval systems. *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information*. pp. 230-239. ACM: New York, USA.

# Building a Distributed Digital Library System Enhancing the Role of Metadata

Gianmaria Silvello
Department of Information Engineering
University of Padua
Via Gradenigo, 6/B
35131 Padova (Italy)
*silvello@dei.unipd.it*

**Abstract**

**In this work we present a methodology that maps a tree data structure in a nested sets organization; the aim of this methodology is to provide an effective and efficient way to manage and exchange descriptive archival metadata in a distributed environment. We consider archival metadata as a basis for our study because of their challenging nature and the management and exchange difficulties they present. Furthermore, we describe a distributed *Digital Library System (DLS)* architecture that would successfully apply the designed methodology. We also introduce future investigations concerning the definition of nested sets organization as an alternative way to manage hierarchical data.**

*Keywords: Distributed Digital Library Systems, Digital Archive Systems, OAI-PMH, Nested Sets Model*

## 1. INTRODUCTION

The role of *Digital Library Systems (DLSs)* in collecting, managing, sharing and preserving our cultural heritage is increasingly prominent in several contexts. DLSs are becoming the fundamental tool for pursuing interoperability between different cultural organizations such as libraries, archives and museums. Collecting and managing the resources of these organizations is fundamental for providing wide, distributed and open access to our cultural heritage.

In this wide and heterogeneous scenario, interoperability is the most relevant issue that a DLS has to face. In a distributed environment the first problem is interoperability between different information systems; a DLS must be able to collect resources shared by a wide number of different systems without compromising their autonomy and independence. We consider a wider context that includes resources of different organization types; in this context the interoperability issue is emphasized by two main necessities. These are the designing of a unique access point to several resources widely different in nature and the cooperation between different information systems.

In this work we consider a challenging kind of information resource: archival documents. When archives are considered, interoperability between the archives themselves, between archival resources and between archival and other types of resources must be taken into account. In the work we have been carrying out we have underlined that DLS technologies need to be revisited to be well-suited and successfully applied to the management of archival metadata and digital objects [1, 2]. In this paper we briefly describe the nature of archival resources, we introduce a methodology based on a nested sets organization able to manage and exchange the archival metadata preserving their whole informative power and we present a DLS architecture that enables them to be included in a DLS. Moreover, we introduce possible future investigations concerning the generalization and formalization of the nested sets organization methodology; in particular, we present the issues that must be considered in order to define the nested sets organization as a way of managing hierarchical data.

The paper is organized as follows: in Section 2 we present the background, projects and initiatives that constitute the context in which this work has been carried out. In Section 3 a brief analysis of archive peculiarities is reported. Section 4 presents the nested sets methodology applied to archival descriptive metadata. Section 5 presents an applicative scenario in which the nested sets methodology can be applied; this applicative scenario is a distributed DLS architecture which we defined in order to share and develop advanced services on archival metadata in a distributed environment. In Section 6 we present future investigations concerning the nested sets methodology.

## 2. BACKGROUND

In order to provide wide access to large and broad collections of digital resources and to address interoperability issues, several initiatives have been instituted. The DELOS Network of Excellence on Digital Libraries[1] has proposed and developed a reference model for laying the foundations of digital libraries [3] which takes into account the perspectives and needs of different cultural heritage institutions and provides a coherent view on the main concepts which constitute the universe of digital libraries in order to facilitate co-operation among different systems.

The "European Commission Working Group on Digital Library Interoperability", active from January to June 2007, had the objective of providing recommendations for both a short-term and a long-term strategy towards "the setting up of the *European Digital Library* as a common multilingual access point to Europe's distributed digital cultural heritage including all types of cultural heritage institutions" [11]. In particular, the recipient of these recommendations is the Europeana thematic network[2], which is a project launched in July 2007 with the aim of addressing the interoperability issues among European museums, archives, audio-visual archives and libraries towards the creation of the "European Digital Library".

Interoperability between different systems has been promoted by the *Open Archives Initiative (OAI)*[3] through *Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)* [20], a flexible and lightweight protocol for metadata harvesting, which is becoming the *de-facto* standard in metadata exchange in distributed environments. This protocol permits metadata harvesting between different repositories in a straightforward fashion, in order to create aggregated metadata collections and to enable the creation of advanced services on them. At the same time *Dublin Core (DC)*, a tiny and lightweight metadata format, is becoming the preponderant means for the exchange of information in a wide distributed environment. Indeed, the characteristics of DC have enabled it to address several interoperability problems and it has been chosen as the minimum common denominator in the OAI-PMH environment. Libraries have been using the couple OAI-PMH and DC for a relatively long time with good results [19].

In fact, in order to develop a large *Digital Library (DL)* among all initiatives, two relevant European initiatives are *The European Library* portal[4] and *Digital Repository Infrastructure Vision for European Research (DRIVER)*[5] which both enjoy the benefits of OAI-PMH. *The European Library* is a free service that offers access to the resources of the 48 national libraries of Europe in 20 languages. The goal of The European Library is to create a single access point to all the European national libraries. The European Library project offers a concrete integration possibility based on OAI-PMH, used to collect the catalogue records of national libraries. Furthermore, the TELplus project[6] will form another building block of the European Digital Library and is aimed at strengthening, extending and improving The European Library service. In particular, to contribute to interoperability among different organizations cooperating in The European Library, it aims to improve and enhance the adoption of OAI-PMH as a means of integration.

[1] http://www.delos.info/
[2] http://www.europeana.eu/
[3] http://www.openarchives.org/
[4] http://www.theeuropeanlibrary.org/
[5] http://www.driver-repository.eu/
[6] http://www.theeuropeanlibrary.org/portal/organisation/cooperation/telplus/

DRIVER is a European project whose goal is to develop a pan-European Digital Repository Infrastructure by integrating existing individual repositories from European countries and developing a core number of services, including search, data collection, profiling and recommendation [4]. DRIVER emphasizes the implementation of nominal, globally accepted standards in a real-life system, with a focus on metadata exchange, in particular using OAI-PMH. One of the Digital Library application components provided by DRIVER is an OAI-Publisher Service; in this way DRIVER services operate upon the aggregated content of existing institutional OAI repositories.

Nevertheless, in the archive context neither general interoperability efforts nor the adoption of specific solutions such as OAI-PMH are common and widespread; this tends to exclude archival documents from forming a valuable part of the cultural heritage managed by a DLS.

## 3. PECULIARITIES OF ARCHIVES

An archive is the trace of the activities of physical people or juridical organizations in the course of their business. Archives have been preserved because of their continued value over time. Archives have to keep the context in which their documents have been created and the network of relationships among them in order to preserve their informative content and provide understandable and useful information over time. In this way archives are able to preserve the provenance of their documents; the preservation of provenance of digital resources is an important issue currently being investigated by the scientific community [15] and must be considered as a key feature of a DLS; it is through provenance information that authenticity can be demonstrated, and the history of archival documents can be preserved.

*Archival descriptions* have to reflect the peculiarities of the archive, retain all the informative power of a *record* and keep trace of the provenance and original order in which resources have been collected and filed by archival institutions [10]. Indeed, archivists seek to group and describe all the archives created by the same organization together, and call this by its French name “*fonds*”. As indicated in [6], the fonds should be viewed primarily as an “intellectual construct”, the conceptual “whole” that reflects an organic process in which a records creator produces or accumulates series of *records*. In this context, provenance becomes a fundamental principle of archives; the principle of the “*respect des fonds*” which dictates that resources of different origins be kept separate to preserve their context [7] suggests that maintaining provenance leads archivists to evaluate *records* on the basis of the importance of the creator’s mandate and functions, and fosters the use of a *hierarchical method* where the hierarchical structure of the archive expresses the relationships and dependency links among the records of the archive by using what is called the *archival bond*. Archival bonds, and thus relations, are constitutive parts of *archival resources*: if an *archival resource* were taken out from its context and lost its relations, its informative power would also be considerably affected.

Archival descriptive metadata are the foremost digital resources shared by the archives. Indeed, most archival documents are not available in digital form, but they are described and represented by metadata; sometimes archival resources are metadata themselves. In the archival context metadata are called archival descriptive metadata and express the archival descriptions.

The use of metadata allows us to exploit DLS technologies and data exchange protocols and apply them to the archival descriptions. Archival description metadata should meet the following three main requisites:

1. **Context**: archival description metadata have to retain information about context of a given record, such as the relations between records and the production environment, as stated by the *respect des fonds* principle discussed above.
2. **Hierarchy**: archival description metadata have to reflect the archive organization which is described in a multi-leveled fashion.
3. **Variable granularity**: archival description metadata have to facilitate access to the requested items, which may belong to different hierarchical levels, with the desired degree of detail and without requiring the whole hierarchy to be accessed.

The only standard defined for archival descriptive metadata is the *Encoded Archival Description (EAD)* metadata format. EAD reflects the archival structure and holds relations between entities in an archive. In addition, EAD encourages archivists to use collective and multilevel description. On the other hand, EAD allows for several degrees of freedom in tagging practice, which may turn out to be problematic in the automatic processing of EAD files. The EAD permissive data model may undermine the very interoperability it is intended to foster. Moreover, EAD files are heavy and difficult-to-move; it has been underlined [13, 16] that the EAD metadata standard is not well-suited for use in a distributed environment.

## 4. A METHODOLOGY FOR MANAGING AND SHARE EAD FILES

Different solutions have been studied to permit archival descriptive metadata exchange in a distributed environment. The proposed solutions suggest the couple DC and OAI-PMH as the means to enable the sharing of archival descriptive metadata and to map EAD files in shareable metadata format. The solution proposed in [18] suggests mapping an EAD file into many tiny and easy-to-move DC metadata. In this approach every DC metadata record generated contains a link to the original EAD file. With this approach there is a strong dependency with the original EAD file that undermines the exchange possibilities of the DC metadata [17]. The solution we proposed in [9] defines a methodology that joins and exploits the characteristics of OAI-PMH and DC. This methodology enables archive hierarchy to be expressed and meaningful relations between archival entities to be preserved by levering the role of the OAI sets. The main idea is to map the archive hierarchy into a combination of OAI sets and DC metadata records.

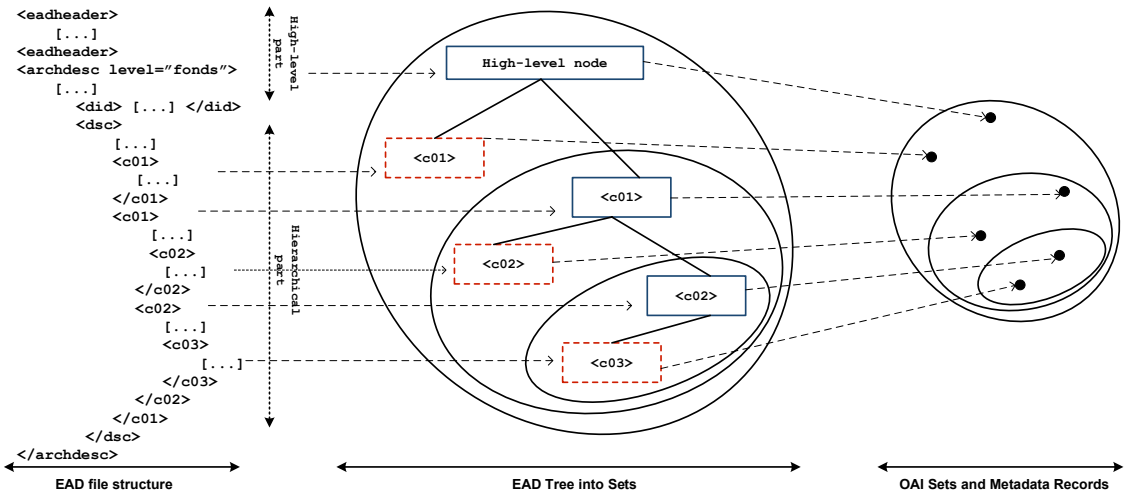



**Figure 1:** Mapping EAD metadata into OAI Sets and DC metadata records.

In Figure 1 we take up the EAD file structure showing how its tree representation can be mapped in a collection of sets[7]. This methodology permits the exchange of archival descriptive metadata in a distributed environment facing interoperability problems and the maintenance of the whole archival informative power of the metadata. We propose a methodology to map the structure of EAD files into several DC metadata records and OAI sets. As far as the mapping of the actual content of EAD items into DC records is concerned, we adopt the mapping proposed by Prom and Habing [18]. We differ from [18] in the way in which the structure of EAD files is translated into OAI sets and DC records. Our methodology[8], shown in Figure 1, is constituted by three main steps:

1. **OAI sets**: creation of an OAI set for each internal node of the tree.

[7] Figure 1 has been reproduced from [9].

[8] To understand this methodology it is worthwhile defining two fundamental characteristics of tree data structure: internal and external nodes. An *internal node* is defined as a node having at least one child, instead an *external node* is defined as a node without children [14].

2. **Metadata set record**: a metadata record for each set constituted in step one must be created; we call these records *metadata set records* because they are built contextually with the OAI set.
3. **Metadata record**: a metadata record for each external node of the tree must be created.

The proposed solution addresses the shortcomings of EAD when it has to be used in a distributed environment and with variable granularity access to the resources. Indeed, EAD items are mapped into different DC metadata records which are shareable metadata, and natively supported by OAI-PMH. Furthermore, context and hierarchy are preserved and expressed in a straightforward manner exploiting the native functionalities of OAI-PMH and DC metadata format. Indeed, the organization into OAI sets reflects archive hierarchy and each metadata record also maintains in its header the membership information which is essential for going up again to related entities and to express contextual information. The proposed solution also addresses variable granularity, indeed a particular archival metadata can be reached without visiting the whole hierarchy.

The proposed nested sets methodology is well-suited to managing and exchanging archival metadata in a distributed environment and it finds an application in the distributed DLS architecture presented in the next section.

## 5. A DISTRIBUTED DIGITAL LIBRARY SYSTEM ARCHITECTURE

A DL aimed at collecting and managing archival resources has to face the complex nature of archives; archive issues can be classified under the interoperability issue and the heterogeneity issue. Interoperability as we have seen is related to the sharing problems of archival resources in a distributed environment; heterogeneity is related to the broad differences between archives and archival resources. Indeed, a DLS has to consider a large number of different archives distributed in a geographical area; each archive exposes a large number of metadata that has to be collected and managed preserving their whole informative power and thus a lot of additional information as well as the metadata themselves.

The constitution of a DLS whose goal is to put archival resources together must take into account the structure and the size of the participating archives. Archives preserve resources that are unique and valuable, thus also small and medium archives need to participate in the system, thus providing important contributions. Usually, independent, private or public archives keep archival metadata without sharing them and this prevents the offering of common advanced services on metadata; one of the goals of a DLS is to provide advanced services on archival metadata. DLSs are service-oriented and can be composed of independent sub-systems that cooperate together to supply the required Digital Library functionalities. Moreover DLSs aim to strengthen integration and interoperability between different systems. The design of a DLS architecture must take into consideration several issues: it has to preserve archive autonomy and gather their metadata to perform advanced services. On the one hand we have to guarantee the bodies maintain archive management autonomy [1]. On the other, we have to build central coordination that has an integrated vision of the archives participating in the system. The added value of this DLS architecture is that it shares metadata exploiting Digital Library advances that can be integrated with and adapted to preexisting systems using different technologies. The result is a scalable, flexible and widely-adoptable architecture for sharing information in a distributed environment [8].

The developed DLS architecture exploits the characteristics of the protocol OAI-PMH based on the distinction between Data and Service Provider and the DC metadata format. The DLS architecture we designed is symmetric in sharing and managing both archival descriptive metadata and authority files treated as metadata too. Indeed, archives act as Data Provider by exposing their descriptive metadata and also as a Service Provider by harvesting the authority files exposed by the central Digital Library. The central Digital Library acts in the same way as a Data Provider furnishing authority files and as Service Provider harvesting archival descriptive metadata. Moreover, the central Digital Library acts as the central authority that constitutes authority files.

The nested sets methodology is a constitutive part of the DLS we propose to share, collect and manage archival resources; this methodology enables several aspects of the interoperability issue to be addressed.

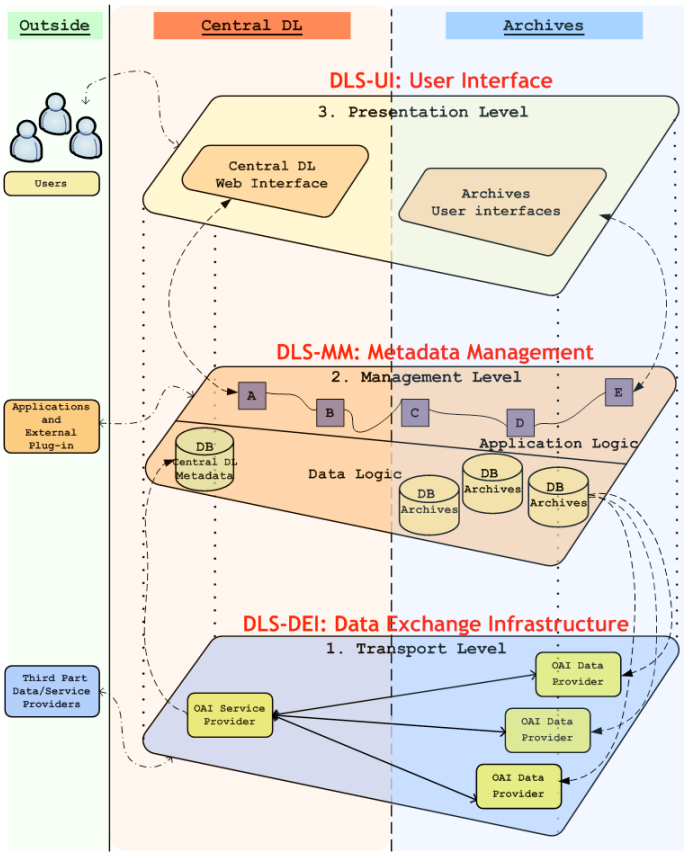



**Figure 2:** DLS Distributed Architecture

The DLS is developed as a three-layer architecture, composed of the metadata transport layer, the metadata management layer and the presentation layer. The transport layer represents the DLS transport infrastructure based on OAI-PMH. The archives participating in the system act as Data Providers providing archival metadata, whereas the central Digital Library acts as a Service Provider that harvests metadata. As stated above, archival metadata have to retain context and hierarchy information; we addressed this issue thanks to a methodology that combines OAI-PMH sets and DC as stated in the previous section. In order to retain this useful and fundamental information the Service Provider has to harvest not only the metadata but also the whole set organization of the Data Provider. Selective harvesting is an OAI-PMH native procedure and it permits effective metadata harvesting that preserves archival information. In this way the archive organization expressed through sets and metadata is recreated in the Service Provider, thus enabling the realization of advanced services on fully expressive archival metadata. The architecture is open to third party OAI-PMH components that for example can harvest the central Digital Library Service Provider.

At the second layer we find the management level called DLS-MM, which is composed of an Application Logic part and a Data Logic part. By the use of Application Logic we can develop advanced services both on harvested metadata owned by the central Digital Library and on the metadata of the archives. The applications developed for the DLS can be used on central Digital Library metadata index and on archive metadata too; indeed they are independent of the transport infrastructure. Thanks to this organization, adding a third-party service to the system will be almost effortless. The Application Logic works on the metadata managed by the Data Logic composed of a central database and a set of distributed databases owned by the archives. DLS-MM Data

Logic preserves and manages the physical data of the system; so this sub-layer manages archive data and central Digital Library archive data as well.

At the third level we have the presentation layer called DLS-UI constituted by the user interfaces. The system presents two main interfaces: the first is a general-purpose interface dedicated to a generic user-type such as archivists, historical researchers, public administrations or private organizations that will use the advanced services available in the DLS; the second is dedicated to specialized users who through this interface can add, remove or update archival metadata.

## 6. FINAL REMARKS AND FUTURE WORKS

In this work we presented a DLS architecture which is able to share, collect and manage metadata in a distributed environment; lightness and scalability of this architectural solution have been shown. Our studies and the designed solutions have produced original contribution in the DL field by proposing an effective solution to handle heterogeneous digital resources.

The research results presented in this paper have been reached during the first year of the PhD of the author; for the future of this work, a new investigation theme concerns the evaluation of the designed methodology that permits the mapping of a tree into a nested sets organization. In the presented case the methodology based on nested sets has been used to solve a specific applicative problem (manage and exchange EAD files); it would be interesting to generalize the methodology in order to evaluate its possible applications in a wider spectrum of problems. The intuitive graphic representation of a tree as an organization of nested sets was used in [14] to show different ways to represent tree data structure and in [5] to explain an alternative way to solve in SQL language recursive queries over trees. However, nested sets representation of tree data structure has not been formalized and generalized yet; a development of this work envisions the formalization of this methodology that we call: “nested sets model”. The definition of the nested sets model requires the study of the set theory in relation to the considered context. We are expected to define the nested sets model starting from the axiomatic set theory; we will formulate, in the context of nested sets model, some of the properties and operations of sets, such as the Zermelo-Fraenkel axioms and the axiom of choice. [12] states that experience has shown practically all notions used in mathematics can be defined, and their mathematical properties derived, in the set axiomatic system. In this sense, the axiomatic set theory serves as a satisfactory foundation for our model based on nested sets.

The formalization of the model will start from an evaluation based on set theory of the mapping methodology. The next steps will involve a redefinition of the methodology and a formal definition of the nested sets model. Afterwards, we will be able to define the meaning of the basic set operations as union, intersection, difference and symmetric difference in the nested sets model, in order to manipulate tree data structure and enable a new way to perform operations on hierarchical data. Throughout the nested sets model it will be possible to design new possibilities in managing, indexing and accessing hierarchical data such as the XML files or define a more effective method for solving recursive queries in relational databases.

## ACKNOWLEDGEMENTS

The author would like to thank his supervisors Maristella Agosti and Nicola Ferro at the University of Padua, Department of Information Engineering. The study is partially supported by the TELplus Targeted Project for digital libraries, as part of the eContentplus Program of the European Commission (Contract ECP-2006-DILI-510003). The work of Gianmaria Silvello was partially supported by a grant from the Italian Veneto region.

## Bibliography

[1] M. Agosti, N. Ferro, and G. Silvello. An Architecture for Sharing Metadata among Geographically Distributed Archives. In C. Thanos, F. Borri, and L. Candela, editors, *DELOS Conference*, volume 4877 of *Lecture Notes in Computer Science*, pages 56–65. Springer,

Heidelberg, Germany, 2007.

[2] M. Agosti, N. Ferro, and G. Silvello. Proposta metodologica e architetturale per la gestione distribuita e condivisa di collezioni di documenti digitali. *Archivi*, 2(2):49–73, December 2007.

[3] L. Candela, D. Castelli, N. Ferro, G. Koutrika, C. Meghini, P. Pagano, S. Ross, D. Soergel, M. Agosti, M. Dobreva, V. Katifori, and H. Schuldt. *The DELOS Digital Library Reference Model. Foundations for Digital Libraries*. ISTI-CNR at Gruppo ALI, Pisa, Italy, November 2007.

[4] L. Candela, D. Castelli, P. Manghi, and P. Pagano. Enabling Services in Knowledge Infrastructures: The DRIVER Experience. In M. Agosti, F. Esposito, and C. Thanos, editors, *Post-proceedings of the Third Italian Research Conference on Digital Library Systems (IRCDL 2007)*, pages 71–77. ISTI-CNR at Gruppo ALI, Pisa, Italy, November 2007.

[5] J. Celko. *Joe Celko's SQL for Smarties: Advanced SQL Programming*. Morgan Kaufmann, 2000.

[6] T. Cook. The Concept of Archival Fonds and the Post-Custodial Era: Theory, Problems and Solutions. *Archiviaria*, 35:24–37, 1993.

[7] L. Duranti. *Diplomatics: New Uses for an Old Science*. Society of American Archivists and Association of Canadian Archivists in association with Scarecrow Press, 1998.

[8] N. Ferro and G. Silvello. A Distributed Digital Library System Architecture for Archive Metadata. In *4th Italian Research Conference on Digital Libraries (IRCDL 2008)*. In print, 2008.

[9] N. Ferro and G. Silvello. A Methodology for Sharing Archival Descriptive Metadata in a Distributed Environment. In B. Christensen-Dalsgaard et al., editor, *Proc. 12th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2008)*, pages 268–279. Lecture Notes in Computer Science (LNCS) 5173, Springer, Heidelberg, Germany, 2008.

[10] A. J. Gilliland-Swetland. *Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment*. Council on Library and Information Resources, 2000.

[11] S. Gradmann. Interoperability of Digital Libraries: Report on the work of the EC working group on DL interoperability. In *Seminar on Disclosure and Preservation: Fostering European Culture in The Digital Landscape*. National Library of Portugal, Directorate-General of the Portuguese Archives, Lisbon, Portugal, September 2007.

[12] K. Hrbacek and T. Jech. *Introduction to Set Theory*. Marcel Dekker, Inc., New York. New York, 1978.

[13] K. Kiesling. Metadata, Metadata, Everywhere - But Where Is the Hook? *OCLC Systems & Services*, 17(2):84–88, 2001.

[14] D. E. Knuth. *The Art of Computer Programming, third edition*, volume 1. Addison Wesley, 1997.

[15] L. Moreau, P. Groth, S. Miles, J. Vazquez-Salceda, J. Ibbotson, S. Jiang, S. Munroe, O. Rana, A. Schreiber, V. Tan, and L. Varga. The Provenance of Electronic Data. *Commununications of the ACM*, 51(4):52–58, 2008.

[16] C. J. Prom. Does EAD Play Well with Other Metadata Standards? Searching and Retrieving EAD Using the OAI Protocols. *Journal of Archival Organization*, 1(3):51–72, 2002.

[17] C. J. Prom. Reengineering Archival Access Through the OAI Protocols. *Library Hi Tech*, 21(2):199–209, 2003.

[18] C. J. Prom and T. G. Habing. Using the Open Archives Initiative Protocols with EAD. In G. Marchionini and W. Hersch, editors, *Proc. 2nd ACM/IEEE Joint Conference on Digital Libraries, (JCDL 2002)*, pages 171–180. ACM Press, New York, USA, 2002.

[19] C. J. Prom, C. A. Rishel, S. W. Schwartz, and K. J. Fox. A Unified Platform for Archival Description and Access. In E. M. Rasmussen, R. R. Larson, E. Toms, and S. Sugimoto, editors, *Proc. 7th ACM/IEEE Joint Conference on Digital Libraries, (JCDL 2007)*, pages 157–166. ACM Press, New York, USA, 2007.

[20] H. Van de Sompel, C. Lagoze, M. Nelson, and S. Warner. The Open Archives Initiative Protocol for Metadata Harvesting (2nd ed.). Technical report, Open Archive Initiative, p. 24, 2003.

# Testing a Genre-Enabled Application: A Preliminary Assessment

Marina Santini
HATII (University of Glasgow, UK)
*MarinaSantini.MS@gmail.com*

Mark Rosso
North Carolina Central University (USA)
*mrosso@nccu.edu*

**In this paper we would like to contribute to the discussion about genre-enabled applications, currently engaging many genre researchers, by presenting a preliminary assessment of a web add-on devised to augment the result list of general-purpose search engines with genre labels. For this assessment, we use a small collection of web pages manually annotated with genre labels by a large number of web users. This resource is made up of two sets of web pages created by two independent researchers for their own user-based genre studies. This comparison allows us to provide a preliminary view on the genre add-on performance and to highlight some open issues in genre research.**

*Genre-enabled applications, web genre, genre annotation, genre labelling, genre evaluation.*

## 1. INTRODUCTION

Genre is a deeply rooted concept in our civilization. Aristotle's *Poetics* started a long-standing discussion about literary text classification by identifying the underlying conventions that differentiate epic, lyric and drama, and the patterns of form and content that characterize tragedy and comedy. Since then, along the centuries, the interest in the conventions typifying textual production has moved from literary criticism, to modern genre analysis, to library science, to online bookshops (e.g. see the *Browse Genres* link in Amazon[1], Figure 1) and finally to digital genres, and genres on the web, a.k.a. web genres. Regardless this uninterrupted tradition of genre studies and practice, the answer to the core question – *what is genre?* – remains basically open due to the number of dissenting definitions, differing characterizations and multiple uses.

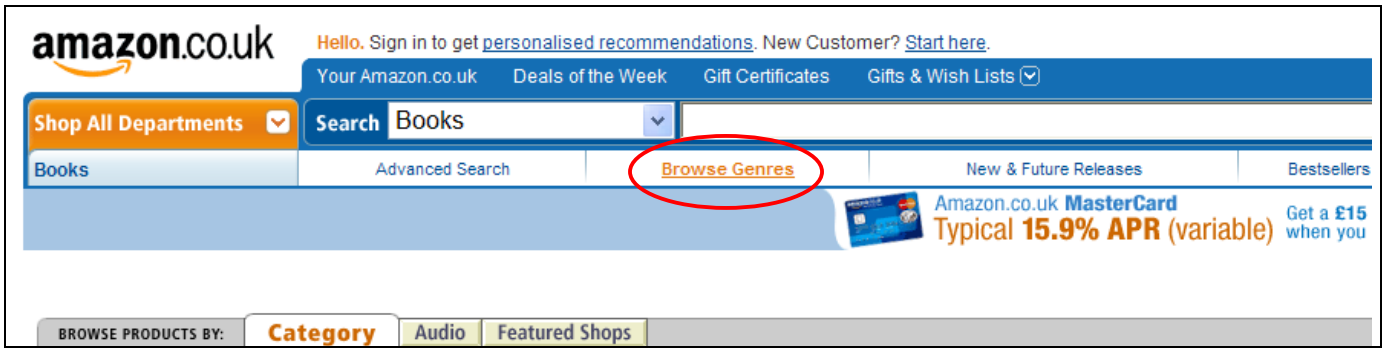



**FIGURE 1.** The *Browse Genres* link in Amazon UK

In this paper, we focus our attention on the genres that can be found on the web and in other digital environments, such as ESHOPS, HOME PAGES, FAQS, ONLINE FORMS, ONLINE TUTORIALS and LIST OF LINKS. The theoretical and empirical studies of digital and web genres are developed within different but related disciplines, such as genre analysis (e.g. Herring et al., 2005; Askehave and Nielsen, 2005), document management (e.g. Orlikowski and Yates, 1994); text technology (e.g. Rehm, 2008), corpus linguistics (e.g. Sharoff, 2007), information retrieval (e.g. Muresan et al., 2006), social network analysis (e.g. Paolillo et al., 2007), web mining (e.g. Mehler and Wegner, 2008), information extraction (e.g. Gupta et al., 2006), automatic summarization (e.g. Seki, 2005), and authorship attribution (e.g. Karlgren and Ericsson, 2007). All the researchers working with genres in these different areas strongly believe that genre is an important classification principle that could help many real-world applications, since genres could be used as filters, as metadata, for indexing, and in any kind of crawlers, agents, or robots that explore or harvest the web or large document collections. Despite its inherent elusiveness, currently there is a high peak of interest in the concept of genre. Many recent academic initiatives have been set up to foster discussion about this concept[2], because – although controversial and difficult to pin down satisfactorily and unanimously – the concept of genre has been proved to be useful to overcome information overload and increase search relevance and usefulness (Freund, 2008a; Freund, 2008b).

[1] See http://www.amazon.co.uk/b/ref=sv_b_1/203-6819510-2539957?ie=UTF8&node=1025612 (accessed 26 August 2008)

[2] For a list, follow the links Reference materials→Genre-Focussed Academic Events in the WEBGENREWIKI <http://purl.org/net/webgenres>.

Genre can be considered a non-topical descriptor that, together with other non-topical descriptors like style or sentiment, may help formulate or refine the information needs expressed in a query. The specificity of genre with respect of other topical and non-topical descriptors lies in its capacity to convey how information is packaged. For instance the INTERVIEW genre indicates that a document contains a dialogue between (usually) two people, one asking questions, the other providing answers.

Genres are based on more or less tight conventions. The identification of these conventions allows people to reconstruct the context in which texts have been produced, together with their purpose and function. In a word, genre is a contextual factor that can be derived from the documents themselves rather than from external human-computer interactions, like clickthroughs or eye tracking. There would be no problem in retrieving genres if all the documents were annotated with genre labels, or contained genre labels in the title, headings and in the meta-content, or if they could be unambiguously correlated with a limited set of topics. If this was the case, genre labels could be simply treated as "terms", and genres could be unfailingly retrieved by current retrieval models. Unfortunately, the scenario is more complex, especially on the web, where genre colonization and genre contamination seem to be widespread, since the web is the crossroads of many communities. At present, there are still many documents or web pages that belong to a certain genre, and could be relevant and useful for a search need, but they do not explicitly contain the name of the genre, or their topics are not predictably correlated with predefined genres. Although general-purpose search engines do a good job when the genre of a document is mentioned in the document itself or in the meta-content, there are still problems when this does not occur, since genre labels can hardly be derived by synonym expansion. This is why the Automatic Genre Identification (AGI) is not entirely term-based or topic-dependent. For instance, obvious cues that could turn out to be useful in the automatic identification of the INTERVIEW genre are: a high frequencies of questions, second person personal pronouns, first personal pronouns, verbs like "believe", "think", "assume", or expressions like "in my opinion", without neglecting graphical hints, like the visual differentiation between questions and answers through the use of paragraph spacing, and other typographical cues such as the use of bold.

Concrete attempts to implement genre-sensitive retrieval models have been made recently. More specifically, Luanne Freund (2008) has presented X-Site, a search system designed and implemented to "test the practical value of making use of task-genre relationships in real-life work environment" (Freund, 2008b: 114). A demo of X-Site was shown at SIGIR 2007 (Yeung et al., 2007).

While X-Site has been devised for professionals (namely software engineers) who can exploit the concept of genre to rapidly find information that is task-appropriate, situationally-relevant and mission-critical for their job, WEGA (an acronym that stands for **WE**b **G**enre **A**nalysis[3]), has been developed at the Bauhaus University Weimar by Prof Benno Stein's team (Stein et al., 2008) for the web and for common web users. WEGA is an add-on that superimposes genre labels a few seconds after the result list is returned by a general-purpose search engine, namely Mozilla Firefox.

These recent applications show that genre-enabled systems are feasible and that genre classes can help improve productivity in the workplace (in the case of X-Site) and offer additional hints about the nature of the web pages listed in the search results. However, the incapacity of defining genre unambiguously has serious repercussions on genre classification and genre labelling. In practical terms, issues that genre researchers constantly face are the following:

(i) How can we say that a genre is a genre, and not another textual category like topic, domain, or style? Although valuable attempts to define the boundaries between these neighbouring categories were made by Lee (2001) and Stein and Meyer zu Eissen (2006), we still do not have any practical criterion that can help us share a common view on these categories.

(ii) What are the cognitive, semantic or pragmatic criteria that people follow when creating a document of a certain genre, or classifying a document by genre? Scholars and researchers suggest different answers to this question, thus creating a plethora of genre classes, selected and defined following contrasting criteria and geared towards different aims.

In this paper, we would like to emphasize the importance of using existing resources (whenever possible) for comparison and cross-testing, especially in a field like AGI, where there are no established benchmarks and where every decision is left to subjectivity, from the selection of genre taxonomies to the creation of genre collections. Comparison and cross-testing help establish relations or correlations between different views and approaches, thus creating a more fertile ground for future research.

While X-Site was evaluated by a user study carried out in December 2005 and based on 32 software service consultants (Freund, 2008b: 129-157), WEGA has not yet gone through any user evaluation to date, because, although publicly available, it is still under development. Leaving the final evaluation to WEGA's creators, here we propose a transversal and preliminary assessment of the WEGA add-on in order to provide some insights *along the way*, i.e. while WEGA is still in a pre-final stage, hoping that these can contribute to WEGA final version and, more in general, to the discussion about genre.

We propose the re-use of a ***small*** number of web pages (50) annotated with genre labels by a ***large*** number of web users. This resource is made up of two sets of web pages created by two independent researchers for their own user-based genre studies, more specifically 20 web pages were labelled by 135 people, and 30 web pages were validated by 257 people. This collection is unique and the two sets have never been used beyond the studies they were devised for. Here we will use this collection to have an idea of the extent to which the genre labels returned by WEGA match the judgement of these two samples of web users. This comparison will allow us to provide a preliminary view on the genre add-on performance, and to highlight some open issues in genre research.

[3] WEGA is freely downloadable from the WEBGENREWIKI (follow the links Reference Materials→Genre-Enabled Applications).

The paper is organized as follows: Section 2 briefly describes the user studies in which the two sets were created; Section 3 presents a comparative analysis; finally Section 4 draws some conclusions and outlines viable future directions.

## 2. WEB PAGES LABELLED BY GENRE IN USER STUDIES

Recently, a number of user studies have been carried out in order to understand which genre classes could be useful for web applications, and especially for genre-enabled search engines.
In 2004, Meyer zu Eissen and Stein (2004) carried out a survey through a questionnaire where university students were asked to come up with genres that could meet their information needs. The data were analysed and researchers worked out eight classes that could cover the genres suggested by the users. These eight genre classes have been incorporated in the current implementation of WEGA (see subsection 3.1).
In 2004, Rosso (2008) carried out a series of studies to identify the genre classes that could improve web searches. After the users proposed their own genres, Rosso developed a palette of 18 genres, and validated this palette through a user study with 257 participants classifying 55 web pages.
In 2005, Santini (2008) set up an online study to investigate the level of disagreement in genre labelling. She presented 25 web pages to the users (135 participants) and suggested 21 genre labels, plus two additional labels (*Add a new type* and *I don't know*) to be used by the participants when they were not satisfied with the suggested labels.
The tangible outcome of Rosso's and Santini's studies is represented by two sets of web pages annotated with genre labels by a large number of web users. We wish to use this collection of two sets to explore to what extent the genre classification performed by WEGA corresponds to human genre labelling. Our empirical study is described in the next section.

## 3. WEGA'S PRELIMINARY ASSESSMENT

The rationale of our preliminary assessment is to investigate WEGA performance in classifying the retrieved documents by genre while it is still under development, since, at this stage, a full evaluation would still be premature. This preliminary assessment will also allow us to highlight some problems that currently affect AGI research.
For this small study, we took all the URLs of the web pages annotated by users in Santini's and Rosso's studies, and specified them in Mozilla Firefox browser with WEGA activated. Figure 2 shows this procedure: the URL was typed in the search box (LHS), the results were labelled by genre by WEGA (see coloured flags next to the heading of the snippets). When it was not possible to retrieve the URL, a simple query, based on the web page headings, was typed in the search box (RHS).

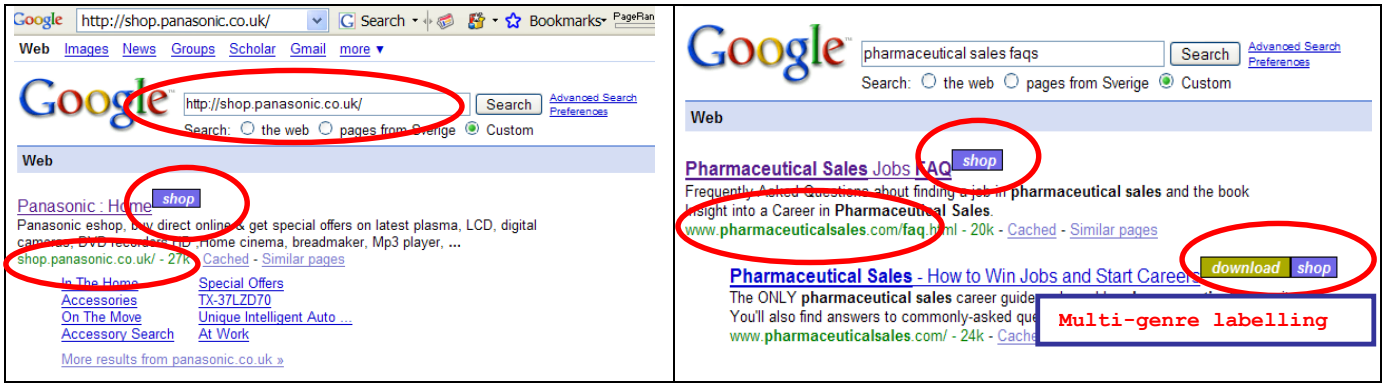



**FIGURE 2:** Searching by URL (LHS) and searching by keywords (RHS)

### 3.1 WEGA

WEGA performs the classification of search results using eight genre classes for German and English web pages. WEGA implementation is described in Stein et al. (2008). The genre classes worked out as described in Meyer zu Eissen and Stein (2004), and implemented in the version of WEGA used in this paper are the following:

1. ARTICLES. Documents with long passages of text, such as research articles, reviews, technical reports, or book chapters.
2. DISCUSSIONS. All pages that provide forums, mailing lists or discussion boards.
3. DOWNLOADS. Pages on which freeware, shareware, demo versions of programs etc. can be downloaded.
4. HELPS. All pages that provide assistance, e.g. Q&A or FAQ pages.
5. LINK LISTS. Documents which consist of link lists for the main part.
6. PORTRAYAL (NON-PRIV). Web appearances of companies, universities, and other public institutions. I. e., home or entry or portal pages, descriptions of organization and mission, annual reports, brochures, contact information, etc.
7. PORTRAYAL (PRIV). Private self-portrayals, i.e. typical private homepages with informal content.
8. SHOP. All kinds of pages whose main purpose is product information or sale.

WEGA is provided also with other non-genre classes, such as “non classifiable”, “unsupported language”, “offline”, etc. WEGA follows a multi-labelling scheme, i.e. the same web page can receive several genre labels (see Figure 2).
We assessed WEGA classification following the mapping proposed in Table 1 and Table 2, and used an assessment scheme similar to the one employed in Freund et al. (2006):

1 = Exact match – when WEGA applies only label(s) assigned by users.
2 = Good match – when WEGA applies all the labels assigned by users plus more.
3 = Fair match – when WEGA includes some of the labels assigned by users.
4 = No match, when WEGA applies no labels assigned by users.

It is worth reminding that WEGA is still under development and a couple of version have already been released to date. The assessment reported here is based on the version issued in March 2008.

### 3.2 SANTINI’s (2008) web pages

The aim of Santini’s study was to explore the need of adopting a multi-labelling genre classification scheme when devising genre-enabled applications, because many experiments in automatic genre classification still focus on only a single label per web page. Her claim was that the single label does not match the view of web users because web pages are often multi-functional and composite. Consequently, when users are forced to select only a single label, they focus on different things, thus generating a large disagreement in genre labelling. Santini’s aim was not to investigate the usefulness of genre for web searches, but to explore the familiarity of web users with labels taken from the web pages themselves. She proposed 21 labels and two escape options (*Add a new type* and *I don’t know*) that the users could use when they were not satisfied with the 21 suggested labels. Labels were suggested in order to reduce the fragmentation that is common with spontaneous users’ terminology (e.g. see the range of variants reported in footnotes 4-14). Her assumption was the following: if users can find an appropriate label in the proposed list, they would gladly use it because this reduces both their cognitive effort and the number of disparate labels (for more details on the sample of participants and the selection of web pages, see Santini, 2008). The labels suggested in Santini’s study are as follows:

1. about page
2. blog (weblog)
3. clog (community log)
4. eshop (online store)
5. email message
6. ezine
7. FAQs
8. home page (academic)
9. home page (corporate)
10. home page (organizational)
11. home page (personal)
12. hotlist
13. howto
14. net advertising (banner)
15. newsletter
16. online frontpage
17. online form
18. online tutorial
19. search page
20. sitemap
21. splash screen
22. *Add a new type*
23. *I don’t know*

In Santini’s study, 135 web users labelled 25 web pages using these labels or their own labels (using the *Add a new type* option), or saying *I don’t know*. Since WEGA palette is coarser-grained than Santini’s labels, we mapped Santini’s labels to WEGA labels following the scheme shown in Table 1. The only direct match was shop-eshop.

**TABLE 1:** Expected matches (Santini’s study)

| **Santini’s (2008)** | **WEGA** |
|---|---|
| eshop | shop |
| about page, corporate home page, organizational home page, | portrait non priv |
| academic home page, personal home page, personal blog | portrait priv |
| sitemap, hotlist | linklist |
| FAQs, tutorials, howtos | help |
| clog, email, newsletter, other blogs | discussion |
| --- | download |
| --- | article |
| online form | *non classifiable* |
| ezine | *non classifiable* |
| net advertising | *non classifiable* |
| newspaper online frontpage | *non classifiable* |
| search page | *non classifiable* |
| splash screen | *non classifiable* |
| *Add a new type* | *(as appropriate)* |
| *I don’t know* | *non classifiable* |

20 of Santini’s 25 original pages could be retrieved and classified by the WEGA add-on The list of the labels assigned by the 135 users and the labels assigned by WEGA is reported in Table 3 (see Appendix). Comparison shows that out of the 20 web pages that could be reached by Mozilla Firefox, there were: **Exact Matches=0; Good Matches=1; Fair Matches=8; No Match=11**.

### 3.3 ROSSO'S (2008) web pages

The genre palette developed by Rosso was the result of three user studies. The first study asked experimental participants to group a set of web pages by genre and assign names and definitions to each genre. In the second study, another set of participants labelled the same set of web pages but their choices were mostly constrained to the 48 labels collected during the first study. Participants were allowed to suggest new labels if none of the other labels were deemed adequate. Rosso used the results of the two studies to create a palette of 18 genres and definitions, derived mostly from user-terminology and preferences. The 18 labels are shown below. A third study with new pages and participants "validated" the palette, achieving over an average of over 70% agreement by the 257 participants for the set of 55 pages (for more details on the sample of participants and the selection of web pages, see Rosso, 2008). The labels suggested in Rosso's study are as follows:

1. article
2. course description
3. course list
4. diary, weblog or blog
5. FAQ/help
6. form
7. forum/interactive discussion
8. index/table of contents/links
9. job listing
10. other instructional materials
11. personal website
12. picture/photo
13. poetry
14. product for sale/shopping
15. search start
16. speech
17. welcome/homepage
18. NONE OF THE ABOVE

Table 2 shows how we expected the web page labels assigned by Rosso's study participants to match up against the labels assigned by the WEGA classifier.

**TABLE 2:** Expected matches (Rosso's study)

| **Rosso (2008)** | **WEGA** |
|---|---|
| article | article |
| FAQ/help/ course description | help |
| forum/interactive discussion archive | discussion |
| index/table of contents/links/course list | linklist |
| personal website/diary, weblog or blog | portrayal priv |
| product for sale/shopping | shop |
| welcome/homepage | portrayal non-priv |
| --- | download |
| NONE of the ABOVE | *non classifiable* |
| speech | *non classifiable* |
| form | *non classifiable* |
| search start | *non classifiable* |
| job listing | *non classifiable* |
| other instructional materials | *non classifiable* |

Only 30 of Rosso's 55 original pages were able to be classified by the WEGA add-on. This was mostly due to pages that no longer exist, or pages that still exist but are no longer in Google's index. We used the same match criteria as with Santini's pages, and the results are reported in Table 4 (see Appendix). Out of the 30 web pages that could be reached by Mozilla Firefox, there were: **Exact Matches=2; Good Matches=0; Fair Matches=11; No Match=17**.

### 3.5 DISCUSSION

Although the user studies were performed between 2004 and 2005 and WEGA web pages were retrieved in March 2008, these web pages had often the same layout and similar content. As already noted by Boese and Howe (2005), "pages in some genres change rarely if at all and can be used in present-day research experiments without requiring an updated version".

Looking at the matches with the two sets of pages combined (i.e. 50 web pages all in all), there were: **Exact Matches= 4%; Good Matches=2%; Fair Matches=38%; No Match=56%** (see Figure 2 for a charted overview of raw counts).

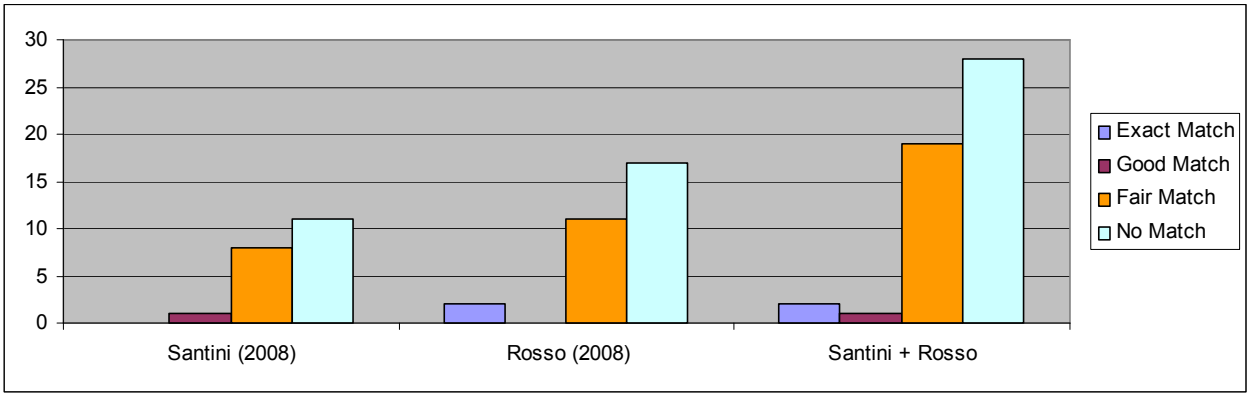



**FIGURE 2:** Overview of the matches (raw counts)

The sum of Exact Matches, Good Matches and Fair Matches is about 45%, which is a promising achievement since genre classification in an open environment like the web is overly difficult for a number of reasons. One reason concerns the distribution of genres on the web. As we do not know how genres are distributed on the web, it is very difficult to approximate web genre population in any computational and statistical model. Another reason is related to feature representativeness, because the relationship between automatically extractable features and web genres is still under exploration. To date, a number of experiments have been carried out to investigate the efficiency and effectiveness of a range of genre features, but always on very small genre collections, containing only a restricted number of genres and a limited number of documents (e.g. cf. Dong et al., 2008; Kim and Ross, 2008; Kanaris and Stamatatos, 2007). In particular, WEGA has been trained and tested on two small but widely used genre corpora, the KI-04 corpus (Meyer zu Eissen and Stein, 2004) and the 7-web-genre collection (Santini, 2007), both containing fewer than 2000 web pages.

An important factor that causes bias in our comparison is the misalignment of Santini's and Rosso's palettes with WEGA palette. Both Santini's and Rosso's palettes contain genres that had no obvious correspondence with WEGA palette. Clearly, WEGA palette is generally at a higher level of abstraction than the other two palettes. In this respect, it would be very helpful for future research to start creating a network of relationships between genres. For example, a new resource could be designed and implemented similar to the hierarchical framework adopted in Wordnet, or following an ontology-like structure. This would permit not only more straightforward comparisons among different genre palettes and different collections, but also a deeper understanding of the cognitive criteria or constraints underlying genre classes.

As far as WEGA is concerned, it seems to be a good idea, for search or browsing purposes, to have more coarse-grained genre classes like HELP including FAQS, TUTORIALS and HOWTOS, or a LINKLIST genre including different kind of listing genres, like SITEMAPS and HOTLISTS. Therefore, a certain level of abstraction in the genre classes presented to web users is very welcomed. However, it is not clear how to distribute more fine-grained genres like BLOGS (usually divided into several subclasses, like PERSONAL BLOGS or NEWS BLOGS, characterised by different purposes and audience) within more general classes. Maybe, BLOGS should have their own place in any genre palette. While BLOGS have a very strong genre identity, classes like PORTRAYAL PRIVATE and PORTRAYAL NON-PRIVATE seem to be rather opaque. For this reason, they have been replaced by other labels, namely PERSONAL HOMEPAGE and NON-PERSONAL HOME PAGE respectively, in the version of WEGA released in April 2008.

Given the small size of the user-labelled web page collection, we cannot draw any final conclusions. As mentioned earlier, a full evaluation (maybe task-based) will probably be carried out by WEGA's creators when this application reaches its completion. With this preliminary assessment, we would like to emphasise that there is still a lot to know about the tradeoffs involved in genre labelling, the variations across different genres, and the cognitive implications in the use of genres (see also Freund et al., 2006).

## 4. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we presented a preliminary assessment of WEGA, a genre add-on. Since no genre test collections or genre benchmarks are currently available, we assessed WEGA performance using 50 web pages annotated with genre labels by the participants to two genre studies. On this small collection, WEGA performance in March 2008 was below 50%. Apparently, there is a wide gap between laboratory tests of genre classifiers and the performance in real-world conditions. Recorded performance measurements for automatic genre classifiers can be higher than 90% (cf. Santini, 2007; Kanaris and Stamatatos, 2007; Dong et al., 2008). If web retrieval by genre is ever to become a widespread reality, this gap must be filled by future research.

Obviously, a reliable and convincing evaluation cannot be based only on a collection of 50 web pages. We consider the assessment described in this paper only a first step towards a larger evaluation. Nonetheless, it is important to stress that this collection, although tiny, is unique because it is labelled by a large number of web users. It is also worth noting that existing genre collections are all annotated with disparate genre criteria, and the annotation is commonly decided on the agreement of a very small number of annotators (at most 4), or decided by the individual researchers themselves. This high degree of subjectivity leads also to diversified genre palettes, and their mapping introduces a bias in subsequent re-use. We suggest that the creation of a genre resource capable of spelling out and encoding the inter-relationships among genres would be useful and would permit a more effective re-utilization of existing resources.

From a methodological viewpoint, WEGA is a web browser add-on. Therefore the retrieval of relevant documents is decided by the underlying search engine on the basis of topical keywords, and WEGA applies genre labels on the search results. A viable and complementary line of genre research would be the integration of topic and genre as combined search criteria, as in Vidulin et al., 2007.

## REFERENCES

Askehave, I. and Nielsen, A. E. (2005). What are the Characteristics of Digital Genres? – Genre Theory from a Multi-modal Perspective. *Proceedings of Hawaii International Conference on System Sciences (HICSS-2005)*.

Belkin, N., Chaleva, I., Cole, M., Li, Y.-L., Liu, L., Liu, Y.-H., Muresan, G., Smith, C., Sun, Y., Yuan, X.-J., Zhang, X.-M. (2005). Rutgers' HARD Track Experiences at TREC 2004. *Proceedings of TREC-2004*.

Boese, E. and Howe, A. (2005). Effects of Web Document Evolution on Genre Classification. *Proceedings of the ACM 14th Conference on Information and Knowledge Management* (CIKM 2005).

Dong, L., Watters, C., Duffy, J. and Shepherd, M. (2008). An Examination of Genre Attributes for Web Page Classification. *Proceedings of Hawaii International Conference on System Sciences (HICSS-2008)*.
Freund, L. (2008a). Situating relevance through task-genre relationships. *Bulletin for the American Society for Information Science and Technoloy*, 34 (5), 23-26.
Freund, L. (2008b). *Exploiting task-document relations in support of information retrieval in the workplace*, doctoral dissertation, Faculty of Information Studies, University of Toronto, Canada <http://faculty.arts.ubc.ca/lfreund/Publications/Freund_Luanne_S_200811_PhD_thesis.pdf>.
Freund, L., Clarke, C.L.A. & Toms, E.G. (2006). Genre classification for IR in the workplace. *Proceedings of Information Interaction in Context* (IIiX 2006).
Gupta, S., Becker, H., Kaiser, G., and Stolfo, S. (2006). Verifying genre-based clustering approach to content extraction. *Proceedings of WWW '06*.
Herring, S., Scheidt, L., Bonus, S. and Wright, E. (2005). Weblogs as a bridging genre. *Information, Technology & People*, 18 (2).
Kanaris, I. and Stamatatos E. (2007). Webpage Genre Identification Using Variable-length Character n-grams. *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence*.
Karlgren, J. and Eriksson, G. (2007). Authors, Genre, and Linguistic Convention. *Proceeding of SIGIR Workshop on Plagiarism Analysis*, 30th International ACM SIGIR Conference, Amsterdam.
Kim Y. and Ross S. (2008). Examining Variations of Prominent Features in Genre Classification. *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS-2008)*.
Lee, D. (2001). Genres, Registers, Text types, Domains, and Styles: Clarifying the concepts and navigating a path through the BNC Jungle. *Language Learning & Technology*, 5(3).
Mehler, A. and Wegner A. (2008). Unifying Content and Structure Learning: A Model of Semi-Supervised Hypertext Zoning. *Abstract Prodeedings of the Processing Text-technological Resources Conference*, Bielefeld University, Germany <http://coli.lili.uni-bielefeld.de/Texttechnologie/Forschergruppe/PTTR/abstracts/Abstract-Mehler-Wegner.pdf>.
Meyer zu Eissen, S. and Stein, B. (2004). Genre Classification of Web Pages: User Study and Feasibility Analysis. In Biundo S., Fruhwirth T. and Palm G. (eds.). *KI 2004: Advances in Artificial Intelligence*, Springer.
Muresan, G., Smith, C., Cole, M., Liu, L. and Belkin, N. (2006). Detecting Document Genre for Personalization in Information Retrieval. *Proceedings of Hawaii International Conference on System Sciences (HICSS-2006)*.
Orlikowski, W. and Yates, J. (1994). Genre Repertoire: The Structuring of Communicative Practices in Organizations. *Administrative Science Quarterly*, 39 (4).
Paolillo, P., Warren, J. and Kunz, B. (2007). Social Network and Genre Emergence in Amateur Flash Multimedia. *Proceedings of Hawaii International Conference on System Sciences (HICSS-2007)*.
Rehm, G. (2008). Hypertext Types and Markup Languages. In Dieter Metzing, Andreas Witt (eds.) *Linguistic Modelling of Information and Markup Languages*, Springer.
Rosso, M. (2008). User-based Identification of Web Genres. *JASIST*, 59(7).
Santini, M. (2007). *Automatic Identification of Genre in Web Pages*. PhD Thesis, University of Brighton, Brighton.
Santini, M. (2008). Zero, single, or multi? genre of web pages through the users' perspective. *IPM*, 44(2).
Seki Y. (2005). Automatic Summarization Focusing on Document Genre and Text Structure. Doctoral Abstract. National Institute of Informatics (NII). Tokyo.
Sharoff, S. (2007) Classifying Web corpora into domain and genre using automatic feature identification. *Proceedings of Web as Corpus Workshop*, Louvain-la-Neuve.
Stein, B. and Meyer zu Eissen, S. (2006) Distinguishing Topic from Genre. *Proc.of I-KNOW 06*.
Stein, B., Meyer zu Eissen, S. and Lipka, N. (2008). Web genre analysis: Use cases, retrieval models, and implementation issues (in preparation).
Vidulin, V., Luštrek, M., Gams, M. (2007). Using genres to improve search engines. *Proceedings of the workshop Towards Genre-enable Search Engines: The Impact of Natural Language*. RANLP-2007.
Yeung, P., Freund, L. and Clarke, C. (2007) X-Site: a workplace search tool for software engineers. System demo presented at the 30th International ACM SIGIR Conference, Amsterdam.

## APPENDIX

**TABLE 3:** Classification of web pages from Santini's (2008) study

| URLs (web page name between brackets, see Appendix in Santini, 2008) | Web Users' Genre Labelling | WEGA Classification | Assessment (1-4) |
|---|---|---|---|
| http://shop.panasonic.co.uk/ (webpage_01) | **ESHOP=88.15%**<br>**NETAD=5.19%** | ***SHOP*** | **2** |
| http://www.satansbarber.co.uk/ (webpage_02) | **PERS. HOMEPAGE=88.89%**<br>**BLOG=7.41%** | ***PORTRAIT PRIV,***<br>***PORTRAIT NON PRIV*** | **3** |

| | | | |
|---|---|---|---|
| http://torvald.aksis.uib.no/corpora/2004-3/0239.htm (webpage_03)<br>**New**: http://gandalf.aksis.uib.no/corpora/2004-3/0239.html | **EMAIL=48.89%**<br>***ADD LABEL=25.19%***[4]<br>**ABOUT PAGE=14.81%**<br>**BLOG=4.44%** | ***PORTRAIT PRIV*** | **3** |
| http://www.nytimes.com/ (webpage_04) | **FRONT PAGE=40.74%**<br>***ADD LABEL=22.96%***[5]<br>**NEWSLETTER=11.11%**<br>**EZINE=8.15%**<br>**ORG. HOMEPAGE=6.67%**<br>**CORP. HOMEPAGE=5.93%** | ***SHOP*** | **4** |
| http://www.dogpile.com/ (webpage_05) | **SEARCH PAGE=83.7%**<br>**ADD LABEL=4.44%**[6] | ***DOWNLOAD*** | **4** |
| http://www.thebritishmuseum.ac.uk/sitemap/sitemap.html (webpage_06)<br>**New**:<br>http://www.britishmuseum.org/about_this_site/site_map.aspx | **SEARCH PAGE=47.41%**<br>**SITEMAP=34.07%**<br>**HOTLIST=7.41%** | ***SHOP***<br>***LINKLIST*** | **3** |
| http://journals.aol.com/brucer5150/AGimpsLife/ (webpage_07) | **BLOG=66.67%**<br>**ABOUT PAGE= 18.52%**<br>**ADD LABEL=7.41%**[7]<br>**DON'T KNOW=4.44%** | ***DISCUSSION***<br>***PORTRAIT NON PRIV*** | **3** |
| http://www.cs.brown.edu/people/ec/ (webpage_08) | **ACAD. HOMEPAGE=58.52%**<br>**PERS. HOMEPAGE=23.70%**<br>**ABOUT PAGE=8.15%**<br>**ADD LABEL=4.44%**[8] | ***DOWNLOAD***<br>***PORTRAIT PRIV*** | **3** |
| http://www.infogistics.com/about.html (webpage_10) | **CORP. HOMEPAGE=69.63%**<br>**ABOUT PAGE=23.70%** | ***DOWNLOAD*** | **4** |
| http://www.intel.com/index.htm?iid=Homepage+Header_UShome& (webpage_11) | **CORP. HOMEPAGE=88.15%** | ***PORTRAIT PRIV*** | **4** |
| http://www.pharmaceuticalsaleshelp.com/faq.php (webpage_12) | **FAQs=83.7%** | ***SHOP*** | **4** |
| http://www.splendidezine.com/ (webpage_13) | **EZINE=60%**<br>**NEWSLETTER=11.85%**<br>**FRONT PAGE=11.11%**<br>**ORG. HOMEPAGE=5.19%** | ***DOWNLOAD***<br>***LINKLIST*** | **4** |
| http://kycares.ky.gov (webpage_14) | **ORG. HOMEPAGE=51.1%**<br>**CORP. HOMEPAGE=9.63%**<br>**ABOUT PAGE=8.89%**<br>**FRONT PAGE=5.93%**<br>**SEARCH PAGE=5.19%** | ***PORTRAIT PRIV*** | **4** |
| http://www.fi.edu/tfi/hotlists/insects.html (webpage_15)<br>**New**: http://www.fi.edu/learn/hotlists/insects.php | **ADD LABEL=23.70%**[9]<br>**HOTLIST=21.48%**<br>**SITEMAP=17.04%**<br>**TUTORIAL=8.15%**<br>**ACAD. HOMEPAGE=5.93%**<br>**SEARCH PAGE=5.19%**<br>**DON'T KNOW=5.19%** | ***PORTRAIT NON PRIV*** | **3** |

[4] The added labels for this page were: *bulletin board, discussion group, discussion list, discussion page, email (within a web-based mailing-list archive), email archive, email discussion list message, email newsgroup archive, forum, forum posting, list serve web posting, listserv, listserv message, listserve message, mailing list, mailing list archive, message board, message board entry, message from newsgroup, newsgroup, online forum, online forum/community interactive page, web forum/discussion list*.

[5] The added labels for this page were: *e-newspaper, electronic newspaper, entry point of a regular newspaper, home page (newspaper), home page (publication), home page newspaper, info webportal, news, news site, newspaper, newspaper front page, on-line newspaper, online magazine, online news source, online news website, online newspaper, online newspaper, periodical front page*.

[6] The added labels for this page were: *home page (search engine), meta-searchengine, portal, search engine, search engine front page*.

[7] The added labels for this page were: *bulletin board, chat page, diary, discussion forum (chat room), entries within a blog, forum, message board, online forum, online journal, someones bull shit* (sic).

[8] The added labels for this page were: *academic's personal home page, contact, online cv, organizational sub-link, personal page on academic institution website*.

[9] The added labels for this page were: *academic document, catalog, classification page, contents page, database listing, encyclopaedia* (sic), *expert information perhaps, index, index of links, index page, information page, itemization page, knowledge directory entry, link list, links page, menu page, navigation page, online encyclopedia, online reference, online table of contents, online textbook, primary navigation tool, reference, reference page, select from list, table of contents, topic indices*.

| | | | |
|---|---|---|---|
| | **ABOUT PAGE=4.44%** | | |
| http://faculty.plattsburgh.edu/nancy.allen/aok.htm (webpage_18) | **ORG. HOMEPAGE=38.52%**<br>**ABOUT PAGE=16.30%**<br>**NEWSLETTER= 10.37%**<br>**DON'T KNOW=10.37%**<br>**HOTLIST=5.93%**<br>**ADD LABEL=5.93%**[10] | ***DOWNLOAD*** | **4** |
| http://wt.xpilot.org/publications/linux/howtos/cd-writing/html/ (webpage_20)<br>**New**: http://tldp.org/HOWTO/CD-Writing-HOWTO-4.html | **HOWTO=54.07%**<br>**TUTORIAL=22.22%**<br>**FAQs=19.26%** | ***DOWNLOAD*** | **4** |
| http://www.citidex.net/896.htm (webpage_21)<br>**New**: http://www.citidex.com/ | **SEARCH PAGE=57%**<br>**ONLINE FORM=9.63**<br>**ADD LABEL=6.67%**[11]<br>**DON'T KNOW= 4.44%**<br>**ESHOP=4.44%** | ***SHOP*** | **3** |
| http://www.intap.net/~drw/cpp / (webpage_22)<br>**New**:http://www.intap.net/~drw/cpp/cpp03_02.htm | **TUTORIAL=65.19%**<br>**HOWTO=21.48%**<br>**ADD LABEL=5.93%**[12] | ***ARTICLE*** | **4** |
| http://www.oceanoptics.com/products/ach.asp (webpage_23)<br>**New**:http://www.oceanoptics.com/Products/74ach.asp | **ADD LABEL=26.67**[13]<br>**ABOUT PAGE=20.74%**<br>**ESHOP=14.81%**<br>**DON'T KNOW=10.37%**<br>**TUTORIAL=6.67%** | ***SHOP***<br>***ARTICLE*** | **3** |
| http://www.lotekk.net/index.php?page=moz&sub=splash (webpage_24) | **SPLASH SCREEN=45.19%**<br>**DON'T KNOW=17.78%**<br>**ADD LABEL=15.56%**[14]<br>**NET AD=4.44%** | ***SHOP***<br>***HELP*** | **4** |

**TABLE 4:** Classification of web pages from Rosso's (2008) study

| URLs | Web Users' Genre Labelling | WEGA Classification | Assessment (1-4) |
|---|---|---|---|
| http://themis.law.ualr.edu:81/ | **INDEX=50%**<br>**SEARCH START=41%** | ***SHOP***<br>***DISCUSSION*** | **4** |
| http://www.matsci.ucdavis.edu/ | **WELC./HOMEPAGE=90%** | ***DOWNLOAD*** | **4** |
| http://www.hnet.uci.edu/mposter/syllabi/readings/viruses.html | **ARTICLE=89%** | ***ARTICLE*** | **1** |
| http://otto.cmr.fsu.edu/~kelley_r/justtonnetz.htm<br>**New**:<br>http://www.robertkelleyphd.com/justtonnetz.htm | **ARTICLE=50%**<br>**PERS. WEBSITE=22%** | ***PORTRAIT PRIV***<br>***ARTICLE*** | **1** |
| http://www.cs.wpi.edu/Research/aidg/CS540/aid.html | **COURSE DESCR.=93%** | ***PORTRAIT PRIV***<br>***ARTICLE*** | **4** |
| http://blogs.law.harvard.edu/ethan/ | **DIARY, BLOG=95%** | ***ARTICLE*** | **4** |

[10] The added labels for this page were: *calendar page, content page, events page, listings page, mixed, organisational page (not home), results page, search result page*.

[11] The added labels for this page were: *entry within the yellow pages, online directory, online service, results page, search result, search results page, specific search page (regional), yellow pages*.

[12] The added labels for this page were: *detailed information, instruction manual, manual, page/section of software documentation, reference, reference page, technical resource/ nerdy/geeky, technical test*.

[13] The added labels for this page were: *(technical) product information page, content page, content page (corporate), corporate web site content, information, information page, normal webpage, online product information, product catalogue, product documentation, product info, product information, product information page - technical specification of a product, product manual, product specification page, product specification sheet, product/ service info (details) page, specification sheet - tech info page, sub page of an online store3, tech spec, tech specs, technical description, technical document, technical documentation, technical documentation/product description, technical information about a product, technical information page, technical instructions, technical product description, technical spec, technical specification, technical specification/product description, technical specifications, technical specifications document, technical specs, webpage*.

[14] The added labels for this page were: *browser, browser loader, dialog box, flash, flash page, flash website, game, game site, loading message, loading page, loading prompt, org loading page, placeholder for a flash app, software download, splash screen, tick-tock page (see comment), wait page, web application*.

| | | | |
|---|---|---|---|
| **New**: http://www.ethanzuckerman.com/blog/ | | | |
| http://helpdesk.wisc.edu/page.php?id=2836 | **FAQ/HELP=76%** | ***DOWNLOAD*** | **4** |
| http://ls.berkeley.edu/mail/webnet/2004/0046.html | **FORUM/INT. DISC.=85%** | ***HELP PORTRAIT PRIV*** | **4** |
| http://www.cropsci.uiuc.edu/faculty/long/ | **PERS. WEBSITE=63%<br>NONE of the ABOVE=19%** | ***DOWNLOAD ARTICLE*** | **4** |
| http://www.english.uiuc.edu/maps/poets/s_z/cdwright/burt.html | **ARTICLE=72%<br>POETRY=49%** | ***ARTICLE LINKLIST*** | **3** |
| http://books.nap.edu/books/0309072786/html/20.html | **SHOPPING=33%<br>ARTICLE=31%** | ***ARTICLE*** | **3** |
| http://www.unt.edu/untpress/titles/davisrod.htm | **SHOPPING=83%<br>ARTICLE=10%** | ***ARTICLE PORTRAIT PRIV*** | **3** |
| http://pharmacy.ucsf.edu/alumni/address/4/<br>**New**:<br>http://pharmacy.ucsf.edu/facultyandstaff/address/4/ | **ARTICLE=28%<br>NONE of the ABOVE=25%<br>SPEECH=18%<br>SEARCH START=16%** | ***PORTRAIT PRIV*** | **4** |
| http://asucd.ucdavis.edu/organizations/other/mar/<br>**New**:<br>http://www.ucdmaar.org/ | **WELC./HOMEPAGE=75%** | ***SHOP*** | **4** |
| http://www.uphs.upenn.edu/pahedu/library/ | **WELC./HOMEPAGE=60%** | ***PORTRAIT PRIV PORTRAIT NON PRIV*** | **3** |
| http://www.biosci.ohio-state.edu/ | **WELC./HOMEPAGE=93%** | ***PORTRAIT PRIV PORTRAIT NON PRIV*** | **3** |
| http://www.math.ucsd.edu/~williams/bandwidth/kwfluid.html | **ARTICLE=63%<br>NONE of the ABOVE=20%** | ***PORTRAIT NON PRIV*** | **4** |
| http://www.med.unc.edu/alcohol/prevention/quiz/quiz.html | **OTHER INSTRUCT.=48%<br>FAQ/HELP=14%<br>NONE of the ABOVE=14%<br>FORM=10%<br>ARTICLE=9%** | ***LINKLIST ARTICLE*** | **3** |
| http://iitc.tamu.edu/1998and2000/lessons/lesson20.html | **OTHER INSTRUCT.= 78%<br>COURSE DESCR.=15%** | ***PORTRAIT PRIV LINKLIST*** | **4** |
| http://home.case.edu/~mss42/2003/11/matchcom-connecting-people-until-they.html<br>**New**:<br>http://liquidschwartz.wordpress.com/2003/11/09/matchcom-connecting-people-until-they-die/ | **DIARY, BLOG=86%** | ***PORTRAIT NON PRIV SHOP*** | **4** |
| http://undergrad-catalog.buffalo.edu/coursedescriptions/index.php?frm_abbr=HIS&frm_num=161 | **COURSE LIST=82%<br>COURSE DESCR.= 12%** | ***HELP ARTICLE*** | **3** |
| http://www.sunysb.edu/philosophy/new/courses/curr_grad_courses.html | **COURSE LIST=67%<br>COURSE DESCR.=30%** | ***ARTICLE LINKLIST*** | **3** |
| http://www.su.edu/conservatory/scon/Courses/MUCH.IDC | **COURSE LIST=80%** | ***LINKLIST PORTRAIT PRIV*** | **3** |
| http://mason.gmu.edu/~lrockwoo/Sample%20Exam%204.htm | **OTHER INSTRUCT.= 81%<br>FORM=12%** | ***ARTICLE*** | **4** |
| http://www.kennesaw.edu/communication/outofstowninternships.shtml | **JOB LISTING=91%** | ***PORTRAIT PRIV*** | **4** |
| http://www.brook.edu/comm/events/20040310iraq.htm | **ARTICLE=51%<br>SPEECH=30%** | ***ARTICLE PORTRAIT PRIV*** | **3** |
| http://web.princeton.edu/sites/chapel/112303.htm<br>**New**:<br>http://web.princeton.edu/sites/chapel/Sermon%20Files/2003_sermons/112303.htm | **SPEECH=67%<br>ARTICLE=20%** | ***ARTICLE PORTRAIT NON PRIV*** | **3** |
| http://www.sfsu.edu/~jtolson/vgarden/1996/garden96.htm | **PERS. WEBSITE=66%** | ***LINKLIST POTRAIT NON PRIV*** | **4** |
| http://all.successcenter.ohio-state.edu/epl259-su2001/module-03-forms/self-survey-3-1.asp<br>New: http://all.successcenter.ohio-state.edu/tmtnm/tmtnm.htm | **FORM=62%<br>OTHER INSTRUCT.=19%<br>NONE of the ABOVE=16%** | ***ARTICLE DISCUSSION*** | **4** |
| http://web.cornell.edu/redesign/blog/index.php?p=32 | **BLOG=75%** | ***ARTICLE*** | **4** |

# Children's information retrieval: beyond examining search strategies and interfaces

Hanna Jochmann-Mannak
Human Media Interaction, Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente,
PO Box 217, 7500 AE Enschede, The Netherlands
*h.e.mannak@utwente.nl*

Theo Huibers
Human Media Interaction, Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente,
PO Box 217, 7500 AE Enschede, The Netherlands
*t.huibers@utwente.nl*

Ted Sanders
Utrecht Institute of Linguistics UiL OTS, Utrecht University,
Trans 10, 3512 JK Utrecht, The Netherlands
*ted.sanders@let.uu.nl*

**The study of children's information retrieval is still for the greater part untouched territory. Meanwhile, children can become lost in the digital information world, because they are confronted with search interfaces, both designed by and for adults. Most current research on children's information retrieval focuses on examining children's search performance on existing search interfaces to determine what kind of interfaces are suitable for children's search behaviour. However, to discover the true nature of children's search behaviour, we state that research has to go beyond examining search strategies used with existing search interfaces by examining children's cognitive processes during information-seeking. A paradigm of children's information retrieval should provide an overview of all the components beyond search interfaces and search strategies that are part of children's information retrieval process. Better understanding of the nature of children's search behaviour can help adults design interfaces and information retrieval systems that both support children's natural search strategies and help them find their way in the digital information world.**

*Information retrieval, search strategies, search interface, information need, conceptualizing, querying.*

## 1. INTRODUCTION

Children's access to the information world is increasingly shifting from the physical library or classroom to the digital world. Every day, more children have access to the internet. Since most children nowadays grow up using computers, they seem to manage working with them better than the average adult. However, can they find relevant information in this giant information world as easily as we all might think they can? Most studies on web usability are focused on adult information-seekers. These studies report all kind of problems adults experience during information-seeking and they offer guidelines how to design user-friendly websites. The study reported in this paper focuses on children's information-seeking and discusses children's search strategies and problems, and research directions to examine how to support children's search behaviour in digital environments.

Most of the information-seeking problems experienced by children are due to the fact that search interfaces are designed by adults. Therefore, design tends to be based on adult search experiences. Unsurprisingly, search strategies required to find information are also based on adults' experience. This causes problems for children, because children are different from adults in many ways: they have other needs than adults and their cognitive, social, physical and emotional development has not yet reached the adolescent formal operational stage of development (Piaget and Inhelder, 1969, in Cooper, 2005*)*. The most obvious differences between children and adults in information-seeking behaviour, relate to interaction style (e.g. children scroll less than adults), navigation style (e.g. adult navigation style is more systematic than child navigation style), relevance (e.g. children use different relevance criteria than adults) and mind set (e.g. children have different concepts and categories in mind than adults). To help children in effective and efficient information-seeking, it is important to know how to give them access to the information world in ways consistent with their learning, cognitive development and curriculum.

The Netherlands Public Library Association (VOB) is aware of the importance of research on children's access to the digital information world. That is why the VOB started a research program to investigate children's search behaviour called 'The digital youth library'. The study reported in this paper is an initial exploration in the domain of children's information retrieval and the study is focused on children of 10 through 12 years which are not yet mature in their use of the internet.

Research on children's information retrieval mostly focuses on testing children's performance and examining their search strategies on a given interface. If a child's performance is, for example, more effective and efficient with a particular search tool than with another search tool, researchers may conclude that this search tool is suitable for children's search behaviour. However, to examine children's natural search behaviour, we believe that research beyond interface is needed by examining 'the black box' of children's information-seeking. That means we have to examine children's cognitive processes when they are searching for information and determine what kind of concepts and categories they have in mind, such as images, shapes, feelings, or genres. It is also important to examine at what level of abstraction children develop concepts.

As a basis for our research, we present a paradigm of children's information retrieval in Section 2, consisting of the components that model the process of a child searching for information after it has been given a particular search task. Search strategies and search interfaces are two important components of this paradigm, but also other components will be described that might even be more important in research on children's information retrieval, such as children's conceptualization and query matching with children's queries.

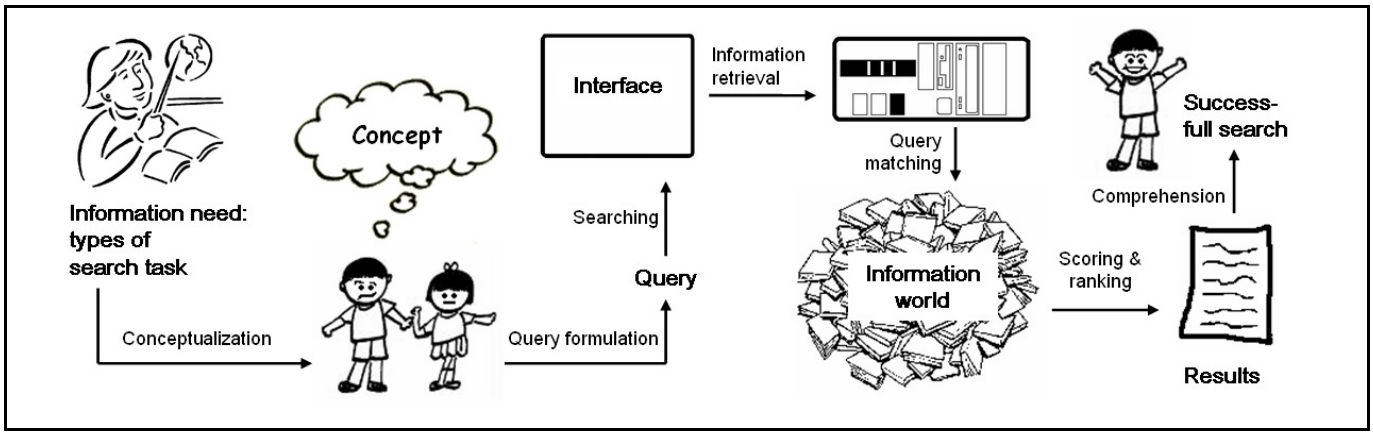
Section 3 discusses current research on children's search behaviour in more detail. We will present research methods and research findings concerning different search strategies and interfaces. Also found difficulties that children come across during information-seeking will be reviewed. Finally, we will discuss that current research on children's information retrieval does not expose the most important problems children encounter during information-seeking. We will discuss what kind of research we believe is needed to discover the nature of children's search behaviour and the real information-seeking problems children cope with, concerning conceptualization and query formulation.

## 2. A CHILDREN'S INFORMATION RETRIEVAL PARADIGM

The domain of children's information retrieval is not limited to searching or browsing on existing search interfaces. First, the child must have a particular search task to formulate a query, for example: what kind of food do most small birds eat? Next, the child has to conceptualize this information need, for example, by displaying an image of a sparrow in his head. After formulating a query, e.g. 'bird food', and feeding this query in an educational website for children, an information retrieval system will try to match this query with relevant documents in the information world.

In the following section every component of the paradigm, as displayed in Figure 1, will be discussed in terms of possible variants and the different effects that variants of a component can have on other components. For example, children of different ages will have different information needs and a child faced with an assigned, fact-driven search task will adopt a different search strategy than a child working on a self-directed search task. All these different variants can be subject to research on children's information retrieval, so as to achieve a better understanding of the nature of children's search behaviour.

**FIGURE 1: A CHILDREN'S INFORMATION RETRIEVAL PARADIGM**



### 2.1 Who are 'they' and what modulations are there?

Children's information retrieval is not just about a child searching for information. First of all, 'a child' is a very broad term. What 'groups' of children are we talking about and do the components in the information retrieval paradigm change by different characteristics such as age, gender, reading skills, computer experience and cognitive developmental stages? To compare different 'groups' of children, it is important to group children with the same characteristics, so that found effects can be associated with the differences in that particular characteristic.

### 2.2 What do 'they' want?

What information are children looking for? In other words, what is a child's information need? What kinds of question do they have? Are these questions mostly self-directed, or externally imposed? Are the questions fact-based or research-based? What is their goal for a search task: to explore, to learn, or to be entertained? Does their information need differ a lot from adults' information need? What do differences in their information needs mean for the way they formulate their information needs in a query or for the way in which they approach a search interface? The effect of change in information need on search performance can be examined by comparing different types of search task in information retrieval experiments. For example, a search task is imposed by a teacher in Figure 1.

Some research has already been conducted on examining search processes and search performance from children performing different kind of search tasks. Schacter et al. (1998) compared children's information-seeking on the internet on two tasks: well-defined tasks and ill-defined tasks. Thirty-two children in the age of 10 to 12 years participated in this experiment. The well-defined task was a closed task (i.e. fact-driven) and had a clearly defined goal in which the information necessary to solve the task was specified in the statement of the task. The ill-defined task was open ended (i.e. research-based): it had vague goals, a large number of open constraints requiring resolution, many possible solutions, and no clear directions for when to stop solving the problem. The researchers found that the children searched more effectively on the ill-defined task than on the well-defined one. Well-defined tasks were difficult for children, because they require highly skilled analytic searching strategies. Ill-defined tasks were easier, because there are more potential answers to ill-defined tasks. They concluded that open-ended and loosely defined search tasks are well suited for children's internet searching. On the other hand, for tasks that are well-defined and highly specific, the internet may not be the most efficient resource to assist children with their information need.

Bilal (2000, 2001, 2002) examined children's use of the Yahooligans! web search engine and compared search performance on three kinds of search tasks: fact-based search tasks, research tasks and fully self-generated search tasks. She observed twenty-two children in the age of 12 to 13 years. Her findings were not in line with prior research (Schacter et al., 1998), because in her research children had more difficulty with the open ended, research task than with the closed, fact-based task. Bilal suggests that these opposite findings can be caused by the children's age differences between the two studies and she claims that more research in examining the effect of different search tasks on search performance is needed. The researcher also found that children were more successful on a fully self-generated task than on the assigned tasks. However, she states that this was due to children's satisfaction with the search results' content rather than the nature of the task itself (i.e. self-generated task).

### 2.3 How do children conceptualize their information need?

To formulate their information need in an utterance or query, first, children have to form a concrete concept of this need in their mind. That is why first of all, it is important to know *what* children think when they search for information. Are they aware that they have to formulate their information need in a concrete query or a search strategy? And does this query vary for different sources? In other words, does a child ask a different question to his mother than to the computer? Second, it is important to know *how* children think when they search for information. At what level of abstraction can they form a query? Can they reach the same levels of abstraction as adults, or can adults think in more abstract terms than children? What kind of categories or concepts do they have in mind: strict taxonomies, prototypes, or emotional categories? What is the role of colours, shapes, images or speech? Is this different from categories or concepts in adult minds? Knowing what kind of concepts and categories are in children's minds is important when we aim at designing interfaces suitable for children's search strategies. Research on concepts and categories in the human mind has for instance, been conducted by cognitive development psychologists, such as Murphy and Lassaline (1997). In a more recent study, Cooper (2005) addresses children's cognitive, physical, social and emotional development that has an impact on a child's ability to interact successfully with a digital environment. She discusses cognitive considerations for designing developmentally appropriate digital environments for young children.

### 2.4 How do children form a query; what are their strategies?

Next, the child has to formulate his or her information need in a 'query'. A query is a command for a source or interface to find relevant information to satisfy the child's information need. What kind of search queries do they form: single concepts, multiple concepts, phrases or natural language? The term 'interface' here has a broad understanding; it can be a digital interface, but it can also be a physical interface such as a bookcase in a library or maybe even a father or mother to whom a child asks a question.

After a child knows what the query will be, he has the possibility to feed this query into a search system. What is his or her strategy and what is this strategy influenced by? Does this strategy differ from adult's search strategies? For example, do children prefer to browse by category or do they want to aim at precisely one goal by inserting a keyword search? Does this strategy change with search tasks? Does it change with designs of interface? In Section 3, research on different search strategies such as searching versus browsing will be discussed in more detail.

### 2.5 What type of interfaces exist and how can information be offered through an interface to support children's search behaviour?

Children's search strategies can be strongly influenced by the way the interface of a system is designed. For example, when an interface does not support browsing, because of category absence, the child has to perform keyword search to find relevant information. Different types of browsing tools - such as word clouds or image clouds, taxonomic search trees, text-based or image-based menus, social or graphical metaphors, simultaneous or sequential presentations (paging or scrolling), clustered versus faceted categories, or flat versus hierarchical presentations - can have different effects on search performance. Additionally, the way in which a search interface is designed in terms of page structure, and pictorial or typographical aspects, can have an effect on children's search performance.

Much research has already been conducted to compare children's search performance with different type of browsing tools or user interfaces. Hutchinson et al. (2006) for example compared a flat, broad and shallow presentation with a deep, narrow hierarchy. With a flat presentation (also termed a simultaneous menu), all items are concrete concepts at a single level. With a deep hierarchy, items are categorized under abstract concepts. The researchers found that for those simple tasks that did not require backtracking, users were faster using the hierarchy, but for more complex tasks, users were faster using flat, simultaneous menus.

Finally, the ways in which search interfaces are displayed can also differ. Search interfaces do not necessarily have to be displayed on a personal computer. Tangible solutions (Price et al., 2003; Blackwell et al., 2004) for displaying a search interface, such as digital tabletops, are another possibility to display search interfaces. The interface of digital tabletops is horizontally displayed to facilitate effective collaboration between children that are working together on the same interface. Sluis et al. (2004) designed Read-It: a multimodal, tangible and collaborative tabletop application for children that supports learning to read in a novel way. In their research on this tangible interface with fifteen children in the age of 5 through 7 years, they found that the interface provided various strategies to support the learning process: recall, rehearsal and collaboration with another child. These strategies to support the learning process are found less using normal desktop interfaces.

The most important question is how to design an interface that best supports children's *preferred* searching behaviour. Ultimately, another question to be addressed is if such an interface also provides children's optimum search performance. Research that is focused on the effect of different types of search interface on search performance will be discussed in more detail in Section 3.

### 2.6 How can an information retrieval system handle a query to find relevant documents from within 'the information world' that will satisfy the child's information need?

The search interface serves as a user-friendly and accessible cover for a child to interact with, but actually the child is interacting with the system behind the interface. This search system is termed an Information Retrieval system (IR-system) and such a system runs the query to find relevant information. To find documents matching the query, the system has to index documents from within the information world. How can the system best select relevant documents for children? Can relevance rules for adults, that IR-systems use to decide if an information item is 'about' (i.e. relevant to) another information item (axioms of aboutness, Huibers and Bruza, 1994), also be applied on IR-systems for children? For example, when an IR-system selects 'mushroom soup' as a relevant result for the query 'toadstool', an adult will agree that this search result is relevant. However, a child may not see the relevance in 'mushroom soup' when he is searching for the house of a dwarf: a toad stool. Another question is how a system

can handle a query while coping effectively with spelling errors or finding synonyms? Manning et al. (2008) give a thorough view on all the aspects of and previously conducted research on the domain of information retrieval.

### 2.7 How can an IR-system present relevant documents?

After an IR-system has run a query, it finds relevant results. It is important to examine how these results can be presented best for children: on the same page on which the child is searching (simultaneous) or on a new page (sequential). It is also important to examine which results must be presented first: the most relevant results by scoring and ranking the matching documents, or the documents that are most referred to by others. Further, how the individual results can best be presented for children is important: with or without a short summary of the found document. Another question is how the link labels of the results must be formulated to help children make the optimum choice. Finally, the differences and similarities between adult and child preferences for all these aspects have to be examined. Search performance on these different variants of result presentation must be examined to help decide what works best for children.

### 2.8 What is successful search and what is relevance for the target group?

What is 'the optimum choice', as mentioned above? What is relevant information for a child? What relevance criteria does a child have available to determine if a result is relevant? Are these criteria different from adults' relevance criteria? Can a child determine whether or not a result comes from a reliable source? Does it even bother a child if a document is relevant or not? What factors influence relevance judgements? Do they change during the progress of a search? What is 'successful search' to a child? In other words, what kind of search results will satisfy a child? Maybe some children will be satisfied with a result that is not even relevant. In the case of the 'small bird-example' a child may be satisfied with information about what bird food to hang in the garden to feed the birds, even though this is not the food birds eat in their natural environment.

In research on children's relevance criteria with ten children in the age of 10 to 11 years, Hirsh (1999) found that students were generally able to articulate their reasons for selecting relevant information. Important relevance criteria in her research were topicality, novelty, interest, clarity and completeness. She also found that relevance criteria changed over time while conducting a search task over a couple of weeks. Furthermore, the students in her research did not think to question the source of the information, the qualifications of the author, or the accuracy of the facts. She concludes that students need more instruction in how to search and navigate electronic resources, and how to judge the relevance of results to meet their information needs. Hirsh also believes that children need training in how to evaluate the authority and accuracy of the information they find.

## 3. BACKGROUND: RESEARCH ON CHILDREN'S INFORMATION RETRIEVAL

As mentioned before, most research on children's information retrieval focuses on search strategies on existing search interfaces. In this section research methods and research findings on search strategies that children use to find information will be discussed. Also various types of search interface and research reported on some of them will be discussed.

### 3.1 Research methods for examining children's information retrieval

Research methods used to examine children's search behaviour and search performance vary from quantitative methods such as online monitoring (Borgman, 1995; Druin, 2003) and recording activities in a browser, to qualitative methods such as discussions with focus groups (Borgman, 1995), interviews (Bilal, 2000, 2001, 2002; Borgman, 1995; Hirsh, 1999) online questionnaires (Druin, 2003), or observation of search sessions (Hirsh, 1999). With online monitoring web logs can be analyzed to gain insight into the total amount of visitors, both to websites and to individual web pages. Attitudes towards the search interface can emerge during discussions with focus groups and interviews with individual users.

In most experiments in which different types of browsing tools are compared, search performance is measured by recording the activities in a browser during task performance ((Bilal, 2001; Hutchinson, 2006; Revelle, 2002; Schacter et al. 1998). The recordings can be viewed at a later time to collect both quantitative data, such as search success, search time, efficiency and errors committed, and qualitative data, such as search behaviour and knowledge about navigation. Also, the user's ability to construct a search query with keyword search can be analyzed from these recordings. A disadvantage of this research method is that it does not give insight into the cognitive processes that occur during task performance. Recordings of task performance only show what users actually did, such as mouse movements, filling in a query, or clicking on a hyperlink. With the eye-tracking research method, eye movements of the user during task performance are recorded. Such eye-tracking data can give a researcher knowledge about the way in which an information-seeker processes particular elements in a digital environment (Ehmke et al., 2007; Guan et al., 2006).

### 3.2 Research findings on searching versus browsing

How can children find relevant documents in the enormous amount of information provided by the internet, to meet their needs? There are many ways of making information more accessible. The way Google achieves this is by far the most preferred by adult internet users. Is the Google way of searching also most preferred by children using the internet? If so, does Google also provide optimum search performance when used by children? The Google search tool works on 'keyword search'. The user enters a query and Google returns relevant documents from the web.

> "Browsing the web is an alternative to searching the web by means of a direct keyword search. Browsing is an interactive process of skimming over information and selecting choices. Browsing relies on recognition knowledge and skills, and requires less well-defined search objectives than does keyword searching." (Borgman et al., 1995)

An information-seeker can browse a website when there are systematic categories that can be selected by the user, such as semantic hierarchies, menus or search trees. Browsing relies on *recognition*. On the other hand, keyword search relies on *recall*; the user has to recall a certain term from memory. Recognition imposes less cognitive load than recall, because more knowledge is needed to retrieve terms from memory than simply to recognize offered terms. That is why a general assumption is made by researchers that browsing-oriented search tools, relying on recognition knowledge, are better suited to the abilities and skills of children than are keyword search tools (Borgman et al., 1995). However, in their research existing of four different experiments with thirty-two children per experiment, aged 9 through 12, Borgman et al. did not find any evidence for their hypothesis. This was due to the fact that keyword search in their experiment was made too easy for the participants by providing the children with the relevant subset of keywords known to match the database.

Hutchinson et al. (2006) found that children are capable of using both keyword search and category browsing, but generally prefer and are more successful with category browsing. The participants in their study were twelve children aged 6 through 7, twelve aged 8 through 9, and twelve aged 10 through 12, equally split between boys and girls. They explain this finding in relation to children's 'natural tendency to explore'. Young children tend not to plan out their searches, but simply react to the results they receive from the IR-system. Generally, their search strategies are not analytical and do not aim precisely at one goal. Instead, they make associations while browsing. This is a trial-and-error strategy.

By tracking the web logs of The International Children's Digital Library (ICDL), Druin (2003) found that, of 60,000 unique users between the ICDL's launch in November 2002 and September 2003, approximately 75% of the searches used category search (browsing), 15% used place search (by selecting a place using a world interface) and just over 10% of the searches used keyword search. Bilal (2000) found in her research on the use of the Yahooligans! Web Search Engine that most of the children (she observed twenty-two children in the age of 12 to 13 years) used keyword search. Only 36% of the searches were performed by browsing under subject categories. This finding may have been affected by the type of search task that was given in this research: a fact-driven query that automatically stimulated children to use keyword search instead of browsing the categories. She also found that children were chaotic in their search performance: they switched frequently between types of searching (i.e. keyword search or browsing), they often looped their keyword searches and selected hyperlinks, and they frequently backtracked. These findings suggest that children want to combine different search strategies during one search task.

Bilal and Kirby (2002) also found that children were more chaotic in their search performance than adults. In their research, they compared search behaviour between twenty-two children (aged 12 through 13) and twelve graduate students. Children made more web moves, they looped searches and hyperlinks more often, they backtracked more often, and they deviated more often from a designated target. The researchers concluded that adults adopted a "linear or systematic" browsing style whereas most children had a "loopy" style. They explain that this "loopy" style can be caused by children's lower cognitive recall, because the web imposes memory overload that reduces recall during navigation. They also found that children scrolled result pages less often than adults.

Schacter et al. (1998) found that with both highly specific and vague search tasks, children sought information by using browsing strategies. In their research on children's internet searching on complex problems with thirty-two children in the age of 10 to 12 years, they reported the following.

> "Children are reactive searchers who do not systematically plan or employ elaborated analytic search strategies."

Finally, Revelle et al. (2002) report on the development of a visual search interface to support children in their efforts to find animals in a hierarchical information structure. To examine searching and browsing behaviour, 106 children (aged 5 through 10) participated in an experiment on this visual search interface. The researchers found that:

> "(…) even young children are capable of efficient and accurate searching. With the support of a visual query interface that includes scaffolding for Boolean concepts, children can use a hierarchical structure to perform searches and construct search queries that surpass their previously demonstrated abilities with the use of traditional search techniques."

### 3.3 Children's problems on information-seeking

In research on search behaviour, researchers often find that children experience difficulties while using both searching and browsing tools. These tools do not take into account children's information processing and motor skills (Hutchinson et al., 2006).

#### *3.3.1 Difficulties concerning motor skills*

Concerning motor skills, children can have difficulties using a mouse, because they process information more slowly than adults. The smaller the object to be clicked on, the longer it takes for a child to click on it (Fitts, 1995, in Hutchinson et al., 2005). Second, many children have difficulty with typing. They are not yet capable of typing without looking at the keyboard, termed touch-typing. Instead, they 'hunt and peck' on the keyboard for the correct keys (Borgman, 1995). That is why typing for children often takes a long time and can lead to frustration.

#### *3.3.2 Difficulties with searching and browsing*

Usually, formulating a search query is difficult for children, because they have little knowledge to base 'recall' on (Borgman et al., 1995; Hutchinson, 2005). Besides, for searching relevant documents using keyword search, correct spelling, spacing and punctuation are needed. Children have difficulty with spelling and often make spelling errors (Borgman et al., 1995). That is why an information retrieval system should be able to handle spelling errors, to help children find relevant documents using keyword search. Deciding on a single keyword is also difficult for a child, because children tend to use a full natural language query. Thus, a system should also be able to handle natural language queries to find relevant information. In a comparison study between children and adults, Bilal and Kirby (2002) found that when children employed keyword searching, most of their queries were single or multiple concepts, just like adults. However, adults employed advanced search syntax, while children did not use this syntax.

With category search, children first of all have trouble finding the right category, because they have little domain-knowledge to decide which category is optimum. In addition, problems with browsing tools are mostly the result of a lack of vocabulary knowledge. Children often have difficulties understanding abstract, top-level headings, because their vocabulary knowledge is not yet sufficient to understand such terms (Hutchinson, 2006). Therefore, formulation of headings should be adjusted to children's vocabulary knowledge, using simple, concrete search terms. Also children may not think hierarchically like adults and may have trouble understanding the way in which hierarchically based categories are constructed. Knowing what their understanding of categories is, can therefore be of great value in designing browsing tools. Bar-Ilan and Belous (2007) tried to understand what browsable, hierarchical subject categories children create by conducting a cardsorting experiment with twelve groups of four children in the age of 9 through 11 years. They suggested terms to the children through 61 cards. The children were free to add, delete or change terms. The researchers found that the majority of the category names used by existing directories were acceptable for the children and only a small minority of the terms caused confusion. Finally, often information in browsing systems is alphabetically displayed, requiring good alphabet skills. Many children have problems with alphabetizing and therefore have trouble finding information in such browsing systems (Borgman, 1995).

#### *3.3.3 Difficulties concerning 'the black box'*

Most browsing tools do not consider how children prefer to search and use search criteria which work for adults. Children use different search criteria than adults. For example, they like to search by physical attributes such as images, colours and shapes (Hutchinson et al., 2005). Also children like to search by concrete genres such as animals or sports, or by feelings and emotions such as happy, sad, scary or sweet. It is important to know which kind of search criteria children use in designing browsing tools. However, we do not know exactly what children's search criteria are and what goes on in their minds when they search for information. That is why we call this 'the black box' of children's information retrieval.

## 4. DISCUSSION: BEYOND EXAMINING CHILDREN'S SEARCH STRATEGIES USING EXISTING SEARCH INTERFACES

According to previous research, most of the problems children experience with searching and browsing are due to search interfaces that do not take into account both children's low motor skills and their different approaches to searching and browsing in comparison to adults. This is because most search interfaces are designed by adults and are therefore based on skills and preferences of adults.

However, does research that focuses on the interface really uncover the heart of the matter? Can we support children's information retrieval just by knowing which search tools do and do not work for them? If research reports that children perform better with a particular search tool in comparison to another search tool, can we then conclude that this search tool also provides an approach preferred by children. Does this approach align with their natural search behaviour? What is their natural search behaviour? Does it give optimum results or do more mature search strategies give better results? What factors, other than interface design, can have an influence on children's search behaviour? Do search strategies change according to the type of search task? Finally, are search strategies different for different kind of children?

To examine the nature of children's search behaviour, we have to go beyond examining performance on existing search interfaces by examining 'the black box' of children's information retrieval. For example, testing an interface using the eye-tracking method, can give insight in the way a child processes a particular search interface, but it does not say anything about the cognitive processes such as conceptualization in a child's brain while conducting the search task. How can we find out what kind of categories children would like to select, maybe based on colours, images, shapes or feelings? Cardsorting (Bar-Ilan and Belous, 2007), for example, is a plausible method to discover children's preferred categories. However, the disadvantage of this method is that term suggestions are given to the child in stead of formulated by the child itself. How can we find out what concepts children develop and at what level of abstraction they develop concepts in their brains? In research children should be encouraged to formulate concepts themselves. To determine children's concept levels of abstraction, children can be showed pictures and asked to name them. For example, a child can name a picture of a canary 'animal' (superordinate level), 'bird'(basic level) or 'canary' (subordinate level) (Murphy and Lassaline, 1997).

Another important question in achieving optimum search results for children concerns relevance. Can an IR-system handle children's queries in the same matter as adult queries? In other words, is a document that is relevant to an adult's query also relevant to a child's query? Can relevance rules for adults, that IR-systems use to decide if an information item is 'about' (i.e. relevant to) another information item (axioms of aboutness, Huibers and Bruza, 1994), also be applied on IR-systems for children? In research, relevance of found results of IR-systems based on adult relevance rules, should be judged by children.

We believe that research addressing these questions will give insight in the real problems children experience with information-seeking, so that we can support them in their search for information. In this research, the impact of factors such as children's motor skills, domain knowledge, searching and browsing skills, reading and writing skills, and alphabet skills must be minimized, so that found effects can only be caused by experimental factors and not by differences in these mentioned factors.

## 5. CONCLUSIONS AND FUTURE RESEARCH

Previous research showed that children have trouble finding information using adult search tools. Therefore, an important question in the area of children's information retrieval is how to design a search interface that is suitable for children. Most of the research on this subject focuses on testing children's search performance on existing search interfaces. However, we believe that such research does not address the real problems children cope with during information-seeking concerning conceptualization and query formulation.

In our research program 'The digital youth library' we will conduct experiments to discover what is in 'the black box' by examining children's cognitive processes during information-seeking. Only in that way it can be discovered what kind of search strategies children prefer and if such strategies indeed give optimum results for children, because then their strategy is not conditioned by the existing adult-based search interfaces. Better understanding of children's search behaviour will eventually help adults design interfaces and IR-systems that better support children's natural search strategies.

ACKNOWLEDGMENTS

This study is funded by The Netherlands Public Library Association and is part of their research program called 'The digital youth library'. This program is a cooperation between several Dutch public libraries to develop a digital youth library. For the development of this youth library, input from academic research is required to design a developmentally appropriate digital environment for children.

REFERENCES

[1] Bar-Ilan, J. and Belous, Y. (2007) Children as Architects of Web Directories: An Exploratory Study. *Journal of the American Society for Information Science and Technology*, **58**, 895-907.

[2] Bilal, D. (2000) Children's Use of the Yahooligans! Web Search Engine: I. Cognitive, Physical, and Affective Behaviors on Fact-Based Search Tasks. *Journal of the American Society for Information Science and Technology*, **51**, 646-665.

[3] Bilal, D. (2001) Children's Use of the Yahooligans! Web Search Engine: II. Cognitive and Physical Behaviors on Research Tasks. *Journal of the American Society for Information Science and Technology*, **52**, 118-136.

[4] Bilal, D. (2002) Children's Use of the Yahooligans! Web Search Engine: III. Cognitive and Physical Behaviors on Fully Self-Generated Search Tasks. *Journal of the American Society for Information Science and Technology*, **53**, 1170-1183.

[5] Bilal, D. and Kirby, J. (2002) Differences and similarities in information seeking: children and adults as Web users. *Information Processing and Management*, **38**, 649-670.

[6] Blackwell, A.F., Stringer, M., Toye, E.F. and Rode, J.A. (2004) Tangible Interface for Collaborative Information Retrieval. *Proceedings of CHI 2004*, Vienna, Austria, 24-29 April, pp.~1473-1476.

[7] Borgman, C.L., Hirsh, S.G. and Walter, V.A. (1995) Children's Searching Behavior on Browsing and Keyword Online Catalogs: The Science Library Catalog Project. *Journal of the American Society for Information Science*, **46**, 663-684.

[8] Cooper, L.Z. (2005) Developmentally Appropriate Digital Environments for Young Children. *Library Trends*, **54**, 286-302.

[9] Druin, A. (2005) What Children Can Teach Us: Developing Digital Libraries for Children. *Library Quarterly*, **75**, 20-41.

[10] Ehmke, C. and Wilson, S. (2007) Identifying Web Usability Problems from Eye-Tracking Data. *Proceedings of HCI 2007*, Rio de Janeiro, Brazil, 20 October, British Computer Society.

[11] Guan, Z., Lee, S., Cuddihy, E. and Ramey, J. (2006) The Validity of the Stimulated Retrospective Think-Aloud Method as Measured by Eye Tracking. *Proceedings of CHI '06*, Québec, Canada, 22-27 April, pp.~1253-1262.

[12] Hirsh, S.G. (1999) Children's Relevance Criteria and Information Seeking on Electronic Resources. *Journal of the American Society for Information Science*, **50**, 1265-1283.

[13] Huibers, T.W.C. and Bruza, P. (1994) Situations, a General Framework for Studying Information Retrieval. *Proceedings of the 16th Research Colloquium of the British Computer Society Information Retrieval Specialist Group*, Drymen, Schotland, March, pp.~3-24.

[14] Hutchinson, H., Bederson, B.B. and Druin, A. (2006) The Evolution of the International Children's Digital Library Searching and Browsing Interface. *Proceedings of IDC '06*, Tampere, Finland, 7-9 June, pp.~105-112.

[15] Hutchinson, H., Druin, A., Bederson, B.B., Reuter, K., Rose, A. and Weeks, A.C. (2005) How do I find blue books about dogs? The errors and frustrations of young digital library users. *Proceedings of HCII 2005*, Las Vegas, NV, 22-27 July.

[16] Manning, C.D., Raghavan, P. and Schutze, H. (2008) *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge. (Preliminary draft, printed on January 25, 2008).

[17] Murphy, G.L. and Lassaline, M.E. (1997) Hierarchical Structure in Concepts and the Basic Level of Categorization. In Lamberts, K. and Shanks, D. (eds), *Knowledge, concepts and categories*. Psychology Press, London.

[18] Price, S., Rogers, Y., Scaife, M., Stanton, D. and Neale, H. (2003) Using 'tangibles' to promote novel forms of playful learning. *Interacting with Computers*, **15**, 169-185.

[19] Revelle, G., Druin, Al, Platner, M., Bederson, B., Houcade, J.P. and Sherman, L. (2002) A Visual Search Tool for Early Elementary Science Students. *Journal of Science Education and Technology*, **11**, 49-57.

[20] Schacter, J., Chung, G.K.W.K. and Dorr, A. (1998) Children's Internet Searching on Complex Problems: Performance and Process Analyses. *Journal of the American Society for Information Science*, **49**, 840-849.

[21] Sluis, R.J.W., Weevers, I., van Schijndel, C.H.G.J., Kolos-Mazuryk, L., Fitrianie, S. and Martens, J.B.O.S. (2004) Read-It. Five-to-seven-year-old children learn to read in a tabletop environment. *Proceedings of the IDC '04 Conference on interaction Design and Children: Building A Community*, Maryland, USA, 1-3 June, pp.~73-80.

# Management and analysis of chinese database extracted knowledge

Nadège Guénec[1], Eloise Loubier[2], Ilhème Ghalamallah[2], Bernard Dousset[2]

1- Université Paris Est Marne la Vallée, laboratoire S3IS, Cité Descartes,
Champs sur Marne, 77454 Marne la Vallée Cedex 2
*nadegeguenec@gmail.com*
2- Institut de recherche en informatique IRIT- SIG,
118, route de Narbonne
31062 Toulouse Cedex 9
loubier@irit.fr
*ghalamal@irit.fr*
*dousset@irit.fr*

**China is an arising country, not only economicaly, but also scientifically. Being aware of the day to day evolution of this emerging country implicates to be able to read the local news, in Chinese langage. In this article we propose to use classical data-mining process tools in an original utilization for analyzing raw datas in order to procure knowledge for business intelligence (BI) application. The aim of this method is, not only to process Chinese datas, but also to create Intelligence by the analyze of the evolution over time of the interactions between specific object within the dataset (key-words, authors, affiliation, so on). The behavior of the environment in the analyzed field will thus be clearly legible throught a summarized representation of the raw datas, thus becoming knowledge. This work focus to provide a new theoretical framework technology for the retrieval information and the management of the associated knowledge, in a BI application. In this paper, we show how to use the data-mining tool and clusters analysis methodology to extract knowledge from a Chinese scientific database, without being able to read Chinese characters.**

*Business Intelligence, knowledge extraction, data mining , data warehouse , relational analysis, evolution, China*

## 1. INTRODUCTION

This work lies within the scope of the research orientations of the CNRS Research Program in Competitive Intelligence. It aims to use tools for automatic data processing developed by the French public research, and in the case of this article, Tétralogie (Dousset, 1988), VisuGraph (Loubier, 2007) and Xplor (Ghalamallah, 2007) tools, for an application to new scientific and technical information "territories": the Chinese information.

The knowledge of China's perpetual change dynamic proves to be imponderable and a need for the survival and the competitiveness of a company. Companies and institutions must be provided with the means of deciphering the strategic issues which are profiled in their environment under penalty of seeing their development slowed down on this area. A forward-looking information retrieval and analysis will prove very useful in order to gain a more relevant visibility of the Chinese environment and its behavior. The results will form part of a comprehensive business intelligence (BI) approach within the company and help the decision maker to better understand its environment.

During the second half of the twentieth century, the development of informatics and communication technology has facilitated the transition to the information age, allowing the emergence of a new industry: the one of knowledge. This industry is mainly driven by databases, which are the containers of human knowledge in various fields of knowledge (Dousset, 1988). Since China opening to the international market in the early 1980's, it had to adapt its industry in order to take root in the international trade and become competitive on the global marketplace. The rapid development of the Internet in China, from the second half of 1990, permits the emergence of the industry of kowledge that had, as elsewhere, to be structured. The database industry is one of the most important sector for scientific and technical information and can be used as an indicator to measure the belonging of a country to the information age. The information industry in China matured together with the emergence of the Internet and the impressive development of scientific research in China. China has thousands of databases, which is a source of information largely untapped in the West in the BI processes.

In the context of the strategic scanning, VisuGraph is a tool particularly adapted to the macroscopic analyses. Indeed, it is able to detect the strong signals, the weak signals and tendencies from a corpus of documents collected for a precise subject. The elaborate information results, represents a synthesis obtained by various methods of data analysis and diffused via graphic visualizations. But because of the different strategic analyses that we have already carried with this software, it appeared that the end users of the produced analyses want, in addition to the general and strategic aspect (general knowledge), more precise views on certain points. In order to satisfy their specific needs for more precise information on elements, which they have already identified (competition, new products or processes, potential partners,…) or in order to discover other elements. Many experts and decision makers are demanding for more details while processing the elements that represent their traditional environment. These elements should contain more precise information about key words, the different actors, the prospective partners and markets that they're coveting for.

In addition, in the business intelligence (BI) context (Ghalamallah and al., 2007), the majority of the strategic information comes from relational sources and the relevance of extracted knowledge usually depends on considering data evolution and their interactions, we propose for our macroscopic analyses a computerized decision-making system with perspective to automate the on-line processing of relational information and to propose analysis and navigation tools oriented to business intelligence (BI) (microscopic).
VisuGraph and Xplor two complementary systems provides strategic analyses on corpora of textual information resulting from the most various sources like: on line databases, Cds, the visible and invisible Web, the news, the press, linking sites, intern databases, etc.
In this article, we present à different experimentations of these systems tested on Chinese data.

## 2. PROPOSITION

Business intelligence (BI) tools enable organizations to understand their internal and external environment through the systematic acquisition, collation, analysis, interpretation and exploitation of information. Two classes of intelligence tools are describe define (Carvalho and al., 2001).The first class of tolls is used to manipulate massive operational data and to extract essential business information from them. The second class of tools, sometimes called competitive intelligence (CI) tools, aims at systematically feeding the organizational environment in order to make possible to learn about it and to take better decisions in consequences. CI depends heavily on the collection and analysis of qualitative information.
This article focuses on the second class of tools, where information is mainly gathered from public sources such as the web, databases, CDROM...
Fuld (Fuld, 2000) describes the CI cycle in five steps:

- Planning and direction: this step is related to the identification of questions and decisions that will drive the information gathering phase.
- Published information collection: search of a wide range of sources, from government fillings to journal articles, vendor brochures and advertisements.
- Primary source collection: this step is related to the importance of gathering information from people rather than from published sources.
- Analysis and production: transformation of the collected data into meaningful assessment.
- Report and inform: delivery of critical intelligence in a coherent and convincing manner to corporate decision makers.

Our approach includes the main phases: analysis and production, report and inform which can be automated by using information technologies (Carvalho and al., 2001).Three big steps handle the data processing and their evolution during a given period: the raw data-gathering, the transformation of the raw data into relational information (pre-knowledge), the extraction of knowledge out of pre-knowledge.

### 2.1 Requirements Formulation

The first step of the BI or CI cycle is the expression by the decision maker of his informational need. A first work has been done about the identification of his needs and targets. Mostly, the requirements are irrelevant or unclear. As this paper is more about the treatment of the data we will just define the informational need with the help of key-words. Our aim here is to prove the validity of the approach and to demonstrate that there is a gap between the existing information and the useful information.

### 2.2 Information Collect and Information Processing

Information processing is based on Knowledge Discovery in Databases (K.D.D.). It defined as « the non-trivial process of identifying valid, potentially useful and ultimately understandable patterns in data», (Fayyad and al, 1996).

For this paper, the extracted corpus was created by articles collected for their link with “wheat” from 2004 to 2008 (seven semesters). Data treatment is divided into five steps:

1rst phase: collecting data

2nd phase: filtering data

3rd phase: cleaning and processing data according to constraints imposed by some tools, algorithms or users

4th phase: crossing the data, providing pre-knowledge.

5th phase: Interaction and data visualization (with the tools called VisuGraph and Xplor).

Our approach has been tested on data extracted from various western databases: SCI, Medline, Pascal, etc…Today, we are testing it with a corpus extracted from a Chinese database: VIP. The process is the same as usual with other databases except that we had to adapt the treatment to the Chinese ideograms.
VIP is a commercial database that provides scientific and technical information. The headquarters of this company is based in Chongqing, at the very center of China. Created in 2000, the base has become the largest Chinese commercial database; they inter alia signed a strategic partnership with Google aimed to adapt "Google Scholar" in Chinese version. In 2005, VIP has been linked to the State information networks, news and publications services which enables it to develop a strategic step on the information industry market.

The database contains more than 12 000 Chinese periodicals and has some 17 million items ordered in eight categories: social sciences, natural sciences, engineering, agricultural sciences and agronomy, medicine and health, economy, education, science and technical, information science and documentation. It is daily consulted in China by tens of millions of readers dispatched in more than 5000 structures: universities, schools, research centers, hospitals, businesses centers, etc… Its reputation has grown because of the completeness of the provided information and also because of the minimal cost of the access. The database is online and the launch of a query is free. The full article (PDF format) can be downloaded through payment with a subscriber's account. An article costs an average of 1 euro. In order to remain on the market of Chinese information industry and position itself to face its rival CNKI (public database developed by Qinghua University), VIP has diversified its offer. In addition of being a scientific and technical literature provider, it also records regularly renewed substantive topics: intellectual property, innovation in Korea, etc… recently have also been started on a statistical overview field about the launched query results.

A last point: the tab "english" which was previously proposed has been recently deleted. More and more difficult thus to access to the contained informations… Anyway, the structure of the data is strictly enforced for all articles; a systematic indexation of the articles realize the possibility for querying by crossing fields and the treatment of the data by infometric tools is thus possible. This is an example of of bibliographic presentation of an article:

【题名】人工合成小麦与普通小麦的重组近交系主要品质性状的初步研究（英文）
【作　者】汤永禄 曾云超 杨武云 邹裕春 陈放
【机　构】四川大学生命科学院,四川成都 610064 四川省农业科学院作物研究所,四川成都 610066
【刊　名】麦类作物学报.2007,27(6).-974-981
【ISSN 号】1009-1041
【C N 号】61-1308
【馆藏号】96016B
【关键词】人工合成六倍体小麦 重组近交系 品质参数
【分类号】S512.1
【文　摘】为了解人工合成六倍体小麦（Synthetic hexapliod wheats,SHW）导入对普通小麦品质的影响及其潜在利用价值,2004-2005 年,对重组近交系群体（人工合成六倍体小麦 Syn-CD780×普通小麦品种 CY12）的主要品质指标进行了检测。结果表明,籽粒硬度、籽粒蛋白质含量、降落值、湿面筋含量、吸水率、形成时间和稳定时间等 7 个品质参数的群体平均值都介于两个亲本之间,只有降落值和面团稳定时间 2 个参数的群体平均值高于 Syn-CD780。在 131 个株系中,有 15 个株系的综合品质指标较为突出。非遮雨处理的籽粒硬度和吸水率显著低于遮雨处理,而籽粒蛋白质含量、降落值、湿面筋含量和稳定时间等 4 个品质指标则相反,表明遮雨与否对小麦品质参数的影响较大。Syn-CD780 在小麦品质改良上有一定的潜在利用价值。

| Tags | Meaning |
|---|---|
| 【题　名】 | Title |
| 【作　者】 | Author |
| 【机　构】 | Laboratory |
| 【刊　名】 | Journal name |
| 【ISSN 号】 | International standard number |
| 【C N 号】 | Chinese number |
| 【馆藏号】 | Collection number |
| 【关键词】 | Keywords |
| 【分类号】 | Subject number |
| 【文　摘】 | Abstract |
| 【网　址】 | Web page |

**FIGURE1**: Bibliographic presentation of an article

Information is synthesized in co-occurrence matrices, used in the various modules proposed by Tetralogie [1][8][9][10].

The basic units of analysis are the term, the field (author, keywords, address, date …) and the document. A field is a basic preset beacon for semi-structured data, as for example author, date, addresses, organization. A field, can have just one value (newspaper) or have different values (author, keyword …).

Data can result from the crossing of two fields, sub-fields or groups of fields in order to obtain co-occurrence matrices. For each of these matrices, crossing between two entities reveals the metric value of the bond between them. Whatever the entity type (author, newspaper,…), it is possible to cross three fields simultaneously. To consider the temporal aspect, the third dimension represents time.

Crossings between two entities are carried out over several homogeneous temporal segments (or periods), in order to analyze the changes induced in time like: absolute changes, relative changes, accelerations, implosions, clusters evolution, etc…

The last step is data visualization and data analysis, giving the following results.

The example in figure 1 shows co-occurrence matrices between authors obtained by Tetralogie. Crossing between the same author give his publications count for each time slice (in bold in the tables below). Crossing between two different authors shows shared publications (co-authors).

| PERIOD 1 | | | | |
|---|---|---|---|---|
| | A | B | C | D |
| A | **5** | 1 | 3 | 0 |
| B | 1 | **2** | 0 | 1 |
| C | 3 | 0 | **8** | 0 |
| D | 0 | 1 | 0 | **1** |

| PERIOD 2 | | | | |
|---|---|---|---|---|
| | A | B | C | D |
| A | **2** | 1 | 0 | 1 |
| B | 1 | **7** | 3 | 3 |
| C | 0 | 3 | **4** | 1 |
| D | 1 | 3 | 1 | **7** |

| PERIOD 3 | | | | |
|---|---|---|---|---|
| | A | B | C | D |
| A | **1** | 0 | 0 | 1 |
| B | 0 | **7** | 3 | 1 |
| C | 0 | 3 | **3** | 0 |
| D | 1 | 1 | 0 | **2** |

| PERIOD 4 | | | | |
|---|---|---|---|---|
| | A | B | C | D |
| A | **1** | 1 | 0 | 0 |
| B | 1 | **9** | 3 | 2 |
| C | 0 | 3 | **3** | 0 |
| D | 0 | 2 | 0 | **3** |

**FIGURE 2**: Publication co-occurrence matrices for four authors {A, B, C, D}, obtained by Tetralogie treatment for four periods consecutively.

### 2.3 Analyzes

This analysis is based on the results of tools for the decision but did not detail their principle of functionality. We invite the interested reader to refer to our research work, to justify development and the principle of each of these features (tufte and al, 1983), (Mothe and al, 1998).
We decide to focus our analysis on the development of authors in the field of wheat in China during the last seven years. In this context, we propose two tools to analyze data, called "VisuGraph" and "Xplor."
With these two tools, the main stages of Business Intelligence (BI) are processed. The originality of these tools based on the submission of a complete system to make both the micro (Xplor) and macro (VisuGraph) analysis, offering a global vision and a specific vision according to the needs of decision-maker (Ghalamallah and al, 2007).

#### *2.3.1 Macroscopic analysis*

Data visualization allows providing as much information as synthetics, which are rarely explained in the raw data. Data representation is an excellent vehicle for analysis of complexity of numerous data ( Tufte, 1983) (Tufte, 1980), (Tufte, 1997). Marks's works on display ( Marks and al, 2005) reveal that a graph may be, clearly, more than two hundred nodes while a computer screen can not display more than twenty consecutive lines, resulting from a search engine classic.
Thus, it becomes easier to analyze arcs between summits, as well as different groupings summits. The overall display documents crossed the keywords can reveal information not visible in the raw data.
Work in visual perception has shown that the human being has a unified global configuration of elements or gestalt-perception of a scene, before paying attention to its details ( Myers, 2000). Work of Tufte (Tufte, 1983) and Bertin (Bertin, 1977) have shown how to exploit, in an intuitive or ad hoc way, these characteristics.
To illustrate the potential of this time-based analysis by VisuGraph tool, we have analyzed the pace of dynamics of actual event data with the study of complex networks of Chinese market relationships as they evolve.
Based on temporal co occurrences matrices on Chinese writers, time graph is drawn through VisuGraph tool. In this graph, nodes represent authors and links represent collaboration between two authors. Each period is visualized on this representation. Each author is represented by a histogram which each bar corresponds to the metric value of the author for a specific period (Loubier, 2007). Thus, the first bar histograms corresponds to the first period, the second in the second period, etc…
Each period is likened to a summit characterizing the year (2002, 2003, 2004, 2005, 2006, 2007, 2008). These are placed on the edge of the window display, equidistant from each other. Peaks representing authors are placed strategically, according to their belonging to different periods. Thus, a certain typology is different. More an author is specific to one year, the more it will be located near the landmark symbolizing this year. The nodes located near a landmark are authors who have written for this specific period. A node located between two landmarks symbolizes the author's persistence on two periods (Loubier, 2007).
In order to detect the most important actors, we apply a filter on data. Based on a specific value, given by the user, every node's value under the threshold is not visible. By this way, the graph is more readable.
The authors located in the center of the graph belong to several periods. The authors are persistent, continually working on this area. The authors are located on the outskirts characteristics to one (or two maximum) periods. Under this chart evolves, we applied a filter to retain only the perpetrators of the most important. It notes the presence of a central core, revealing the presence of authors during the eight periods. However, there is also a strong presence of major authors in the field, during 2006 and 2007. It notes that those authors who began to publish in the first year 2002 have become pioneers and persist on other periods. The strongest circle contains

some of the most important actors. These nodes are connected so we can see that the most important actors work together during the different period. We are interested in those authors who are the most important area. These authors are characterized by their large size histograms. This means that these authors have published more during recent years. Moreover, these authors are more connected, which shows their many collaborations. We print only their names so as not to overburden the drawing. We circle these authors in order to better visualize in the graph. We then get the next figure.

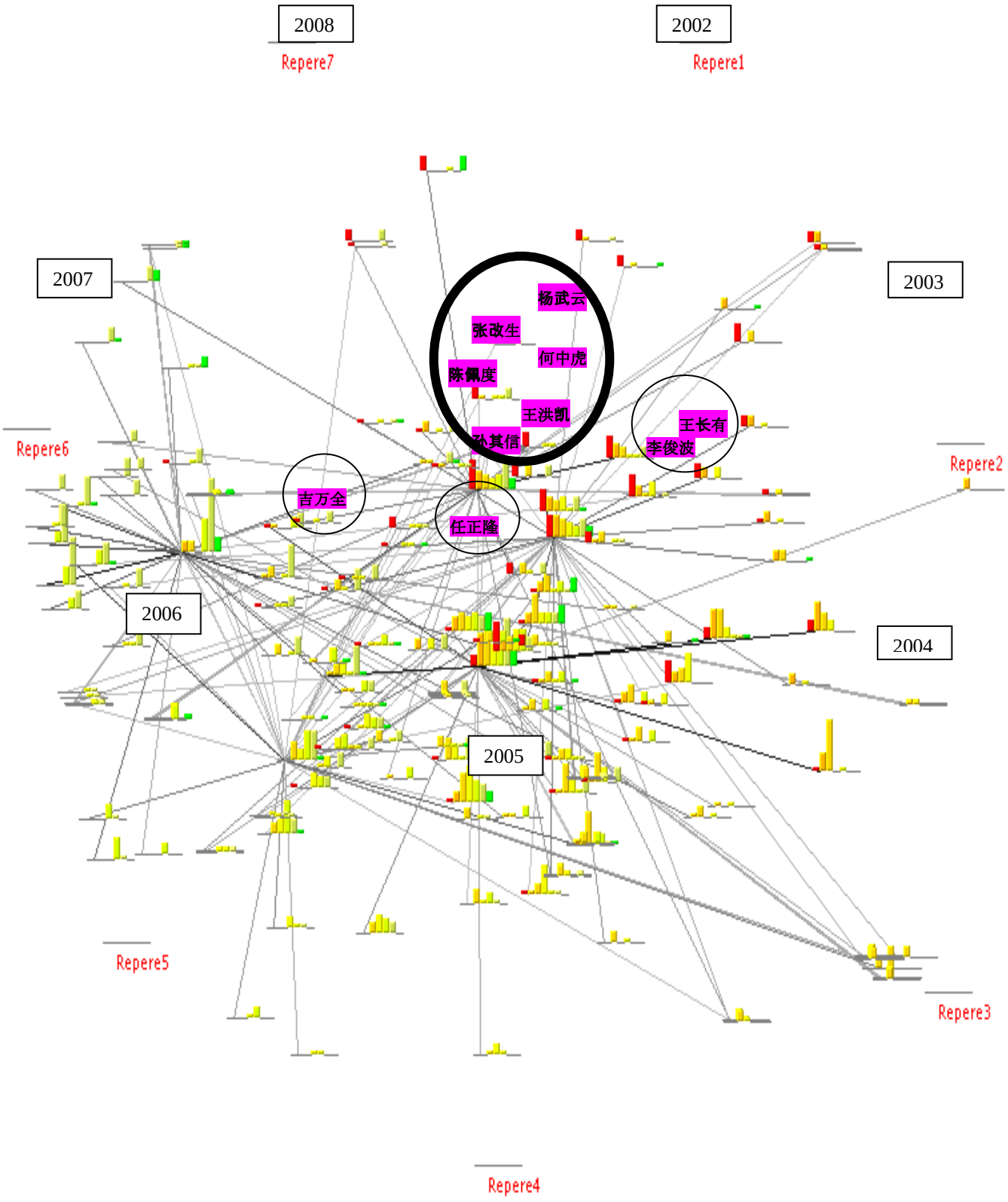



**FIGURE 3**: Data visualization about Chinese authors working from 2002 to 2008.

*.3.2 Microscopic analysis*

Cooccurrences matrix presented before are used for this analysis. We are going to transfer these matrices in the form of BDD to feed the web portal XPLOR. Once the BDD is online, we define different areas of analysis for the user. To complete the study macroscopic achieved with the tool VisuGraph, we developed an indicator of the evolution of the sponsors, through the portal Xplor. To get the evolution of the ten best writers in the existing database on the seven periods. The principle is to zoom on matrices created during the macroscopic analysis. For the display of our results, we offer several types of outputs and graphic output in tabular form. For our experiments on the study of wheat field in china changing authors in this field are represented according to the following figure:

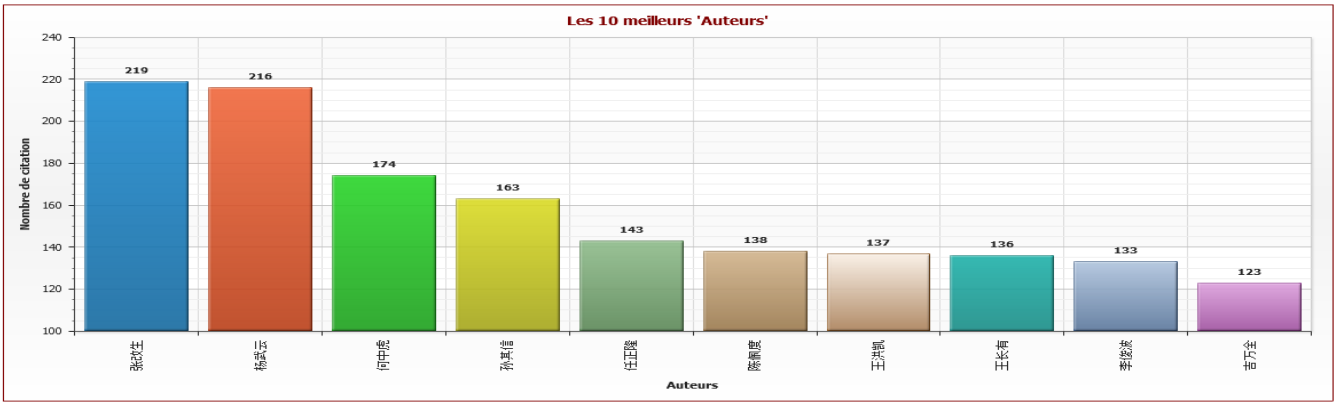



**FIGURE 4***:* Top 10 of authors about wheat between 2002 and 2008.

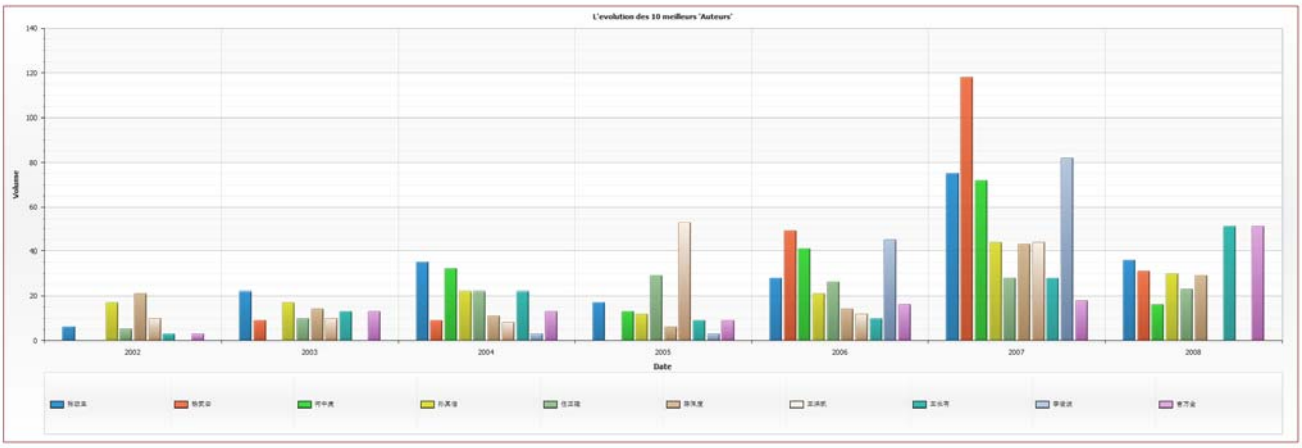



**FIGURE 5***:* Evolution of the ten most published authors about wheat between 2002 and 2008.

We can also view this table in the form cross-table:

| Name | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | Total |
|---|---|---|---|---|---|---|---|---|
| 张改生 | 6 | 22 | 35 | 17 | 28 | 75 | 36 | **219** |
| 杨武云 | 0 | 9 | 9 | 0 | 49 | 118 | 31 | **216** |
| 何中虎 | 0 | 0 | 32 | 13 | 41 | 72 | 16 | **174** |
| 孙其信 | 17 | 17 | 22 | 12 | 21 | 44 | 30 | **163** |
| 任正隆 | 5 | 10 | 22 | 29 | 26 | 28 | 23 | **143** |
| 陈佩度 | 21 | 14 | 11 | 6 | 14 | 43 | 29 | **138** |
| 王洪凯 | 10 | 10 | 8 | 53 | 12 | 44 | 0 | **137** |
| 王长有 | 3 | 13 | 22 | 9 | 10 | 28 | 51 | **136** |
| 李俊波 | 0 | 0 | 3 | 3 | 45 | 82 | 0 | **133** |
| 吉万全 | 3 | 13 | 13 | 9 | 16 | 18 | 51 | **123** |

**FIGURE 5**: Evolution of the ten most published authors about wheat between 2002 and 2008.

## 3. CONCLUSION

The results of the macro and micro analysis are synthesized in the following figure :

| | VisuGraph | Xplor | |
|---|---|---|---|
| **Authors** | «张改生», «何中虎», «杨武云», «陈佩度», «任正隆», «孙其信», «王洪凯», «李俊波», «王长有», … | 张改生 | 219 |
| | | 杨武云 | 216 |
| | | 何中虎 | 174 |
| | | 孙其信 | 163 |
| | | 任正隆 | 143 |
| | | 陈佩度 | 138 |
| | | 王洪凯 | 137 |
| | | 王长有 | 136 |
| | | 李俊波 | 133 |
| | | 吉万全 | 123 |

**TABLE 1**: Synthesis of micro and macro analysis.

We note that the results obtained by VisuGraph (macro) correspond to the results obtained by Xplor (micro). Based on comparable results, these two approaches are intended to complement each other. Indeed, the macro analyzis gives an overview of the evolution of the datas in time and bring out different associations, collaborations and alliances. The most important actors in the field are easily visible but there is no specific classification between them and the others. This is where the micro analysis full fill the macro analysis. Indeed, Xplor brings a more precise analysis to classify the authors according to their importance.

In this paper, we propose an analytical tool dedicated to Business Intelligence to cover all phases of the process of discovery, extraction and data management. In a first step, the data are collected and preprocessed. Then two types of analyses are proposed: a macro analysis (via VisuGraph) and a micro analysis (via Xplor). Both analyses are complementary. The macro analysis gives a comprehensive vision of data analysed, while the micro analysis target in the study to obtain a ranking of the most important.

VisuGraph reveals the detection of the tendencies during time (emerging actors, persistent actors…). This tool makes it possible to easily visualize various alliances between the authors and makes it possible to detect the most important actors (connecting different groups). This tool allows the fast assistance for the decision but does not indicate an importance order (there is no quantitative classification of data). It thus insists on the qualitative one rather than on the quantitative one.

The Xplor tool supplements VisuGraph by the classification and the quantitative valorisation of the data. The most important authors, from a quantitative point of view, will be at the head of classification. However the importance of the authors in terms of alliance will not be proposed. By this way, we do not show that the both tools gives the same results but they complete themselves on the same base. The most important actors detected by the both tools are the same. VisuGraph indicates the temporal characteristics of actors, and the different collaboration between them. Xplor gives quantitative information about these actors.

The presented study focuses on Chinese market and in particular authors in this field. The aim of this paper was almost to show that this tool also can be used to process Chinese extracted corpus. As China is a developping country with strong potential, this work contributes to enlarge the sources for surveying a scientific and technical domain of research by including Chinese data.

Our prospects concerning the two tools would be to assemble in the form of a single tool. Thus we would obtain macro and a micro analysis by the means of one only tool.

With regard to the Chinese analyzes, we could more develop the part “cleaning of data” which is effective for the Western languages but knows limits with regard to the practice of Chinese. It would be necessary to have a complete Chinese dictionary of synonymies.

## 4. REFERENCES

DOUSSET B., BENJAMAA T., Trilogie logiciel d'analyse de données, Conférence sur les systèmes d'informations élaborées : Bibliométrie – Information Stratégique – Veille technologique, 1988.

LOUBIER E., BAHSOUN W., DOUSSET B. "La prise en compte de la dimension temporelle dans la visualisation de données par morphing de graphe". In : Colloque Veille Stratégique Scientifique et Technologique (VSST 2007), Marrakech, 21/10/2007-25/10/2007, IRIT, (support électronique), october 2007b.

GHALAMALLAH I. "L'analyse relationnelle en ligne au service de l'intelligence économique". Dans : Colloque Veille Stratégique Scientifique et Technologique (VSST 2007), Marrakech, octobre 2007.

GHALAMALLAH I., GRIMEH A., DOUSSET B. "Processing data stream by relational analysis". Dans : REVUE MODULAD n°36 (p.67-70), Mai 2007.

CARVALHO, Rodrigo Baroni de & Ferreira, Marta Araújo Tavares (2001) "Using information technology to support knowledge conversion processes" Information Research , vol 7, (2001)).

FULD & Company Inc. (2000) Intelligence software report.

FAYYAD U., Piatetsky-Shapiro G., Smyth P. “The KDD process for extracting useful knowledge from volumes of data”, Communications of the ACM, v.39 n.11, p.27-34, Nov. 1996.

MOTHE J., DOUSSET B. "Analysis of evolutionary trends in astronomical literature using a knowledge discovery system: Tétralogie". Library and Information Services in Astronomy III Conference, LISA III, publié dans ASP Conference Proceedings Series, 1998.

MOTHE J., DKAKI T., DOUSSET B. "Information mining in order to graphically summarize semi-structured documents". Rapport de recherche IRIT/00-22-R, 2000.

MOTHE J., CHRISMENT C., DKAKI T., DOUSSET B., EGRET D. "Information mining: use of the document dimensions to analyse interactively a document set". 23rd BCS European Colloquium on IR Research: ECIR, Darmstadt. BCS IRSG, pp 66-77, 4-6 avril 2001.

TUFTE E. “The visual display of quantitative information”. Graphic Press. Cheshire, p. 198, Connecticut, 1983.

TUFTE E. “Envisioning Information”. Graphics Press, 1990.

TUFTE E. “Visual Explanations”. Graphics Press, 1997.

MARKS L., Hussell J., McMahon T., Luce E. “ActiveGraph : A digital library visualization tool”, Research Library, Los Alamos National Laboratory, USA, Springer-Verlag, 2005.

MYERS D.G. Psychology(6th edition), Worth Publishing, (2000).

BERTIN J. La Graphique et le Traitement Graphique de l'information, Flammarion, (1977).

# Selective Erasers: A Theoretical Framework for Representing Documents Inspired by Quantum Theory

**Alvaro F. Huertas-Rosero**

**Department of Computing Science, University of Glasgow**

**alvaro@dcs.gla.ac.uk**

**The problem of representing text documents within an Information Retrieval system is formulated as an analogy to the problem of representing the quantum states of a physical system. Lexical measurements on text are proposed as a way of representing documents which is akin to physical measurements on quantum states. The representation of the text is only known after measurements have been made, and because the process of measuring may destroy parts of the text, the document is characterised through erasure. These so called “Selective Erasers” provide the basis for representation [1]. During my presentation, the mathematical foundations of such a quantum representation of text will be presented and discussed in the context of indexing and retrieval within a ``quantum like'' Information Retrieval system. The presentation will then outline the various directions for future work using Selective Erasers.**

**Information Retrieval, Quantum Theory**

**Acknowledgements:**

**I would like to thank my supervisors C. J. van Rijsbergen and Leif Azzopardi for their valuable and fruitful advice. Also, I would like to thank the European Commission for supporting my PhD studies through the K-Space Project (FP6-027026), and the Department of Computing Science of the University of Glasgow for the travel support through the Robert’s fund.**

**References**

**[1] Huertas-Rosero, A., Azzopardi, L., van Rijsbergen, C.: "Characterising through erasing: A theoretical framework for representing documents inspired by quantum theory". In: Quantum Interaction 2008, Oxford, U. K., College Publications (2008)**
**[2] van Rijsbergen, C. J "The Geometry of Information Retrieval" Cambridge University Press, 2004**