Informer

Number 25, Winter 2007/8
ISSN 0950-4974

BCS INFORMATION RETRIEVAL

In This Issue

Editorial 1
by Tony Rose

Introduction to the Special Issue 2
by David E. Losada and Juan M. Fernández-Luna

Feature Articles

"Inventing the Future Internet" 3
by Yahoo Research

"The IR Lab at the University of A Coruña" 5
by Javier Parapar and Álvaro Barreiro

"Applying IR in a Contextualized Warehouse" 8
by Juan Manuel Pérez, Rafael Berlanga and María José Aramburu

"Search Oriented Transform of Web Sites" 12
By César Llamas, Pablo de la Fuente and Jesús Vegas

Forthcoming Events 14
Edited by Andy MacFarlane

About Informer

Informer is the quarterly newsletter of the BCS Information Retrieval Specialist Group (IRSG). It is distributed free to all members. The IRSG is free to join via the BCS website (http://irsg.bcs.org/), which provides access to further IR articles, events and resources.

The British Computer Society (BCS) is the industry body for IT professionals. With members in over 100 countries around the world, the BCS is the leading professional and learned society in the field of computers and information systems.

Informer is best read in printed form. Please feel free to circulate this newsletter among your colleagues.

And now for some thing completely different - a "Spanish" Informer.

That's right – this edition is a one-off special, focusing exclusively on the IR scene in Spain. There's a great mix of articles, covering both the latest developments from the academic community and a review of one of the newest and most influential commercial research organisations – Yahoo's Barcelona labs.

All this has been brought together by our guest editors, David Losada and Juan Fernandez Luna. They have provided their own introduction, so I'll keep mine brief. Suffice it to say that if you'd like to see more like this, or even put together your own special issue, then just drop us a line at informer@bcs.org.

In the meantime, all the best for 2008, and enjoy the issue.

Best regards,
Tony

Tony Rose, PhD MBCS CEng
Editor, Informer
Vice chair, IRSG
Email: irsg@bcs.org.uk

Introduction to the Special Issue

by David E. Losada and Juan M. Fernández-Luna

Welcome to a very special Informer issue dedicated to "spanish" IR research! Computer Science research in Spain has been traditionally influenced by areas such as Artificial Intelligence, whose research community is very large. In contrast, Information Retrieval was not a prominent research area within the Spanish map until the late nineties. Nevertheless, the Spanish IR community has been increasingly growing in the last ten years. Many Spanish research teams participate now in international IR events such as ECIR or SIGIR in a yearly basis. Reflecting this grow, the 27th edition of ECIR was even organized in Spain for the first time in 2005.

In this issue, we try to give a small snapshot of some of the research conducted currently by IR Spanish teams. However, we would like to warn you that the list of groups included in this issue is not exhaustive at all (mainly because of size constraints). Many other research groups could have been included here. We simply selected a few representative groups and we expect that the resulting issue sketchs somehow current research trends in "Spanish IR".
The issue starts with an article from the recently created Yahoo! Research team in Barcelona where they present their lab and research lines.
The issue includes also an article from the IRLab of the University of A Coruña, where they present their main research lines and, particularly, NowOnWeb, which is a news web retrieval application developed by the IRLab members. Additionally, the TKBG group from the University Jaume I report on how they have been applying IR in a contextualized warehouse. This is a fruitful research line that ended up in a PhD dissertation presented in 2007. How to transform websites to enhance information access is an important topic of research of the IR & DL group from the University of Valladolid and, in this Informer issue, three researchers from this group explain their recent developments in this direction.
As guest editors, we thank all these groups for their contributions and we hope that IR research in Spain keeps increasing.

Enjoy this issue!

David E. Losada (Univ. Santiago de Compostela) & Juan M. Fernández-Luna (Univ. Granada)

Flag and Bell Pub Crawl

The Flag and Bell is a Tech Pub Crawl held on the first Tuesday of each month.

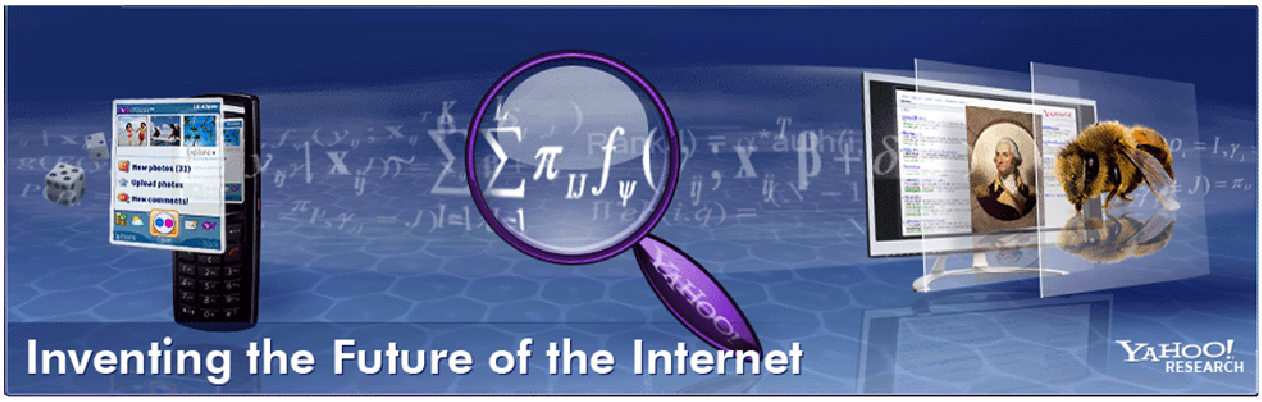
It is a free, networking event for anyone interested in search engines, web technology and the Internet.

We will visit a couple of pubs and spend about 45 minutes at each. All pubs are in walking distance of the first one. All you need to do is listen out for the bell and follow the flag to the next pub.

Date: Tuesday, 5th February 2008, 6pm
Venue:
Exchange Bar (Soho)
32 Gerrard Street
London W1D 6JA

To register: email megan@trexy.com

For more information:
http://flag-and-bell.com/



Yahoo! Research Barcelona:

Yahoo! Research Barcelona[1] was the first research lab opened in 2006 by Yahoo! outside the United States, together with the one in Santiago de Chile. The director of both Yahoo! Research Barcelona and Santiago de Chile is Ricardo Baeza-Yates. This article describes some of the main research areas of interest for these labs.

There are mainly three research groups. One research group is focused on Web search and data mining, including applications of Natural Language Processing (NLP) to Information Retrieval (IR) and mining usage and link patterns. Another group is multimedia IR, with a focus on developing tools for searching and mining audiovisual material. Finally, a group works on parallel and distributed IR developing novel algorithms for dealing with massive datasets under heavy query loads.

Web mining

Statistical information about the usage patterns of users can be an extremely valuable source of data, in particular to learn how to serve the needs of the users better. Usage data on the Web comes mainly in two flavours: access logs and query logs. Access logs are records of the activities of a user on a Web site, while query logs are records of the activities of a user on a search engine. Respecting the privacy of the users is a first concern here, and the privacy of users can be preserved as the focus is on using aggregate information about their collective intelligence. One possible representation of query logs is through graphs, for instance, bipartite graphs whose nodes are queries and documents, and edges connect a query *q* and a document *d*, for instance, whenever a user searching for *q* has clicked on document *d*. This graph can be also interpreted as a query-to-query graph in which novel associations among queries can be discovered. Another application is to attempt query classification exploiting the knowledge about the typical behaviour of users.

Link mining

Links are very important sources of information on the Web and other collections. They have been used successfully in bibliometrics to assert the quality of certain resources, and can be used for ranking Web pages and even, for instance, to predict the medium-term success of an academic paper by observing the citations it receives in the short term. Another application of link mining is to detect deceptive "link farms" that attempt to deceive a search engine's ranking algorithm to receive an undeservedly high ranking.

NLP applications for IR

The aim of this research group is to develop tools that exploit natural language processing and semantic metadata to go beyond current search paradigms. A fundamental step in that direction, is to enrich the logical view the search engine has of the documents being indexed. This can be done by using an statistical taggers and fast dependency parsers, appropriately trained to new and noisy collections. A second step needed is to develop text indexing algorithms which can answer (fast!) queries over the annotated documents. Finally, new ranking algorithms are needed in order to sort linguistically annotated content. This content may be a document or a sentence, as in traditionally search engines, but also an entity (e.g. a

1 http://research.yahoo.com/

person, or an event). For example, we are currently exploring the use of "entity graphs" as an alternative representation of textual information for the task of entity search and browsing.

Beyond these technical challenges, this research group is trying to formalize new tasks in order to bridge the gap between academic research (in NLP, IR and Semantic Web communities) and practical search problems on the web. To this end, the group works with large on-line textual collections (such as Wikipedia, Yahoo! Answers, web advertisements, etc.) and tries to define adequate tasks and performance measures. For example, the group has recently released a snapshot of the English Wikipedia[2] which has been semantically tagged using open source products. As well as the annotated textual data, thy released entity containment graphs and other representations of the data so that people from different communities can experiment with it.

Multimedia IR

The Multimedia group focusses on mining existing knowledge in social multimedia services such as Flickr, Jumpcut, and Delicious, to enable semi-automated annotation of images and video fragments and to deploy new retrieval techniques for multimedia services at large. Furthermore, the group participates in the SEMEDIA project, which is partially supported by the European Community under the Information Society Technologies (IST) priority of the 6th Framework Programme for R&D.

Yahoo!'s social media properties drive on the interaction of users with various types of media, such as photos, videos, and text. When the right incentives are created for the user, this interaction, e.g. human annotation, provides a wealth of information that is commonly referred to as user-generated content (UGC). The research therefore focusses on how to create these incentives for the user, and on how to deploy the obtained knowledge in a broad range of multimedia applications.

Distributed IR

One current research goal of this group is to build a prototype of a realistic large-scale distributed search engine. To reach this goal, they have been working on problems such as caching and query routing. Caching exploits the temporal locality of queries: many queries appear frequently, and many others have one or more terms in common. Query routing is relevant when a number of sites cooperate to provide the search service, and the system has to route queries to the most appropriate site(s). A Web search system can benefit from informed solutions that extract information from different sources to make decisions upon their configurations; Web mining plays a crucial role in such solutions.

30th European Conference on Information Retrieval (ECIR)

30th March - 3rd April
University of Glasgow

ECIR is the annual conference of the BCS-IRSG. It is the major European forum for the presentation of new research results in the field of Information Retrieval.

The conference includes a programme of tutorials and workshops, as well as an Industry Day of presentations and discussion dedicated to the interests and needs of Information Retrieval practitioners.

Registration:
For registration and further details, please visit: http://ecir2008.dcs.gla.ac.uk/

2 http://www.yr-bcn.es/semanticWikipedia

The Information Retrieval Lab at the University of A Coruña

By Javier Parapar and Álvaro Barreiro

The Information Retrieval Lab is affiliated to the Department of Computer Science of the University of A Coruña (code *G000494* in the University catalogue). The group has been researching in basic issues of Information Retrieval for more than ten years.

Álvaro Barreiro leads a team of doctors and young Ph.D. students that are all involved in different research projects, funded by regional and national governments, related with the study and development of new IR techniques and solutions. Several results were published along this time in the main journals and conferences of the field, all they can be found in the group's web page.

More precisely the research was centred in retrieval models for IR, efficiency issues like document identifier reassignment or static pruning, effectiveness in the retrieval of relevant sentences, summarisation, clustering, etc. We also maintain research lines about:

- Basis issues: retrieval models, crawling, indexing, etc.
- Evaluation: methodologies, statistical significance, etc.
- Document classification: Bayes, k-NN, SVM, etc.
- Multimedia IR: Video IR an Audio IR.

The research carried out aimed the group, from the last two years, to consider the possibilities of technology transfer to local companies, as result of that, the group is currently developing real operational systems to solve some of the companies needs. Since our research results and publications are available in the IRLab web page, we want to devote this report to present two of our IR products that are successful examples of the current effort in the area of transfer of technology: NowOnWeb[3], a information retrieval system for the management of on-line news, and the Coruña Corpus Tool, for the management of linguistic corpus, developed in collaboration with the Muste Group of the English Department of our University.

NowOnWeb

From the last decades, the amount of news sites available on-line has suffered a great increase. The communication departments of the big companies and institutions always want to be up-to-date from the relevant news.

To cover this new source of information the manual and traditional techniques are not suitable any more. So we decided to develop a system to help to the press offices. We called it NowOnWeb and we had several contacts with local companies to exploit it commercially.

NowOnWeb can be defined as a NewsIR system that deals with the on-line news sources. It supplies with an effective and efficient approach to show news articles, about a specific topic, to the user in a comfortable way.

The system was component-based designed and comprise a crawler to obtain the web pages, an indexer to maintain the incremental index whit a temporal window, a news recognition and extraction module that enables the dynamic adding of sources, a news grouping component that uses novelty and redundancy detection approaches, and a summariser, among others.

3 http://nowonweb.dc.fi.udc.es

Figure 1: Snapshot of the NewsIR system.

As commercial product NowOnWeb offers:

- A web service that serves the news about the user information needs.
- A flexible and adaptable application to the needs of the users and their areas.
- An efficient and scalable product with effective results.
- A component based software with modules that allows the reuse in different applications.
- A great product to the press areas of the companies and institutions that also can be adapted to other fields as technology surveillance or vertical search.

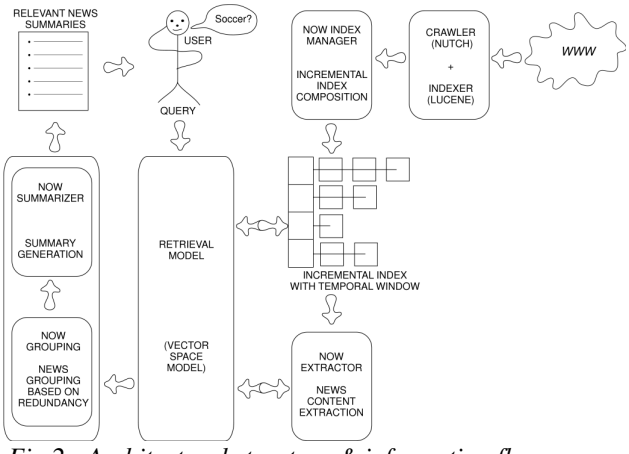


Fig 2: Architectural structure & information flow.

During the analysis and design of the system three main research topics were faced and the results were captured in components:

Now Extractor: An article recognition and extraction algorithm was developed with two objectives: to be extremely efficient in order to be operational and to achieve good results. We got an algorithm based on domain specific heuristics over the articles structure that achieved both objectives. (Further details: J. Parapar & Á. Barreiro in CAEPIA 2007)

Now Grouping: A news redundancy filtering algorithm. It was developed with the aim of avoid the overload of the user reading several times redundant articles. The technique was developed based on traditional approaches in the filtering field. As result of our method, news redundancy groups are created were each group involves redundant articles about the same topic, and each group has a representative that is the article chosen to be shown to the user, while the others are only referenced as links. (Further details: J. Parapar, Á. Barreiro, J. M. Casanova in EUROCAST 2007).

Now Summariser: The most relevant article for each redundancy group is summarised by the system. For this task we applied an approach based on extraction of relevant sentences to the query in retrieval time. We also have developed other several strategies to this problem balancing efficiency and effectiveness in the construction of summaries (Further details: J. Parapar in FDIA 2007).

NowOnWeb is always in developing phase because we use it as a research platform. In this sense we are approaching, among others, architectural system improvements, efficient query logging storage and mining for personalisation issues, and enabling video news support.

The Coruña Corpus Tool

As previously mentioned this is a development carried out by the IRLab in collaboration with the English Department. Indeed the application came up because the need of the Muste Group of having a system to manage and exploit its linguistic corpus, which is still in compiling process. (Further details: J.

Parapar & I. Moskowich-Spiegel in *Procesamiento del Lenguaje Natural* Vol. 39)

The Coruña Corpus contains English scientific texts produced between 1650 and 1900. For each discipline two texts per decade are selected containing around 10,000 words, excluding tables, figures, formulas and graphs. (Further details: I. Moskowich-Spiegel & J. Parapar in AEDEAN 2007). In order to retrieve information from the compiled data, the Coruña Corpus Tool (CCT) was created and it is currently in testing phase. The objective is help linguists to extract and condense valuable information for their research. But the application was not designed tied to the Coruña Corpus and it supports any xml-formatted corpus being, in this sense, an application that could be widely used.

The texts were conveniently formatted following the TEI (Text Encoding Initiative) standard. Several data was tagged in order to allow multi-field search.

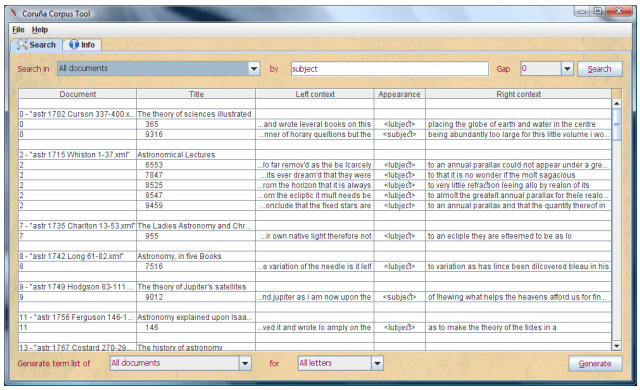


Figure 3: Coruña Corpus Tool snapshot.

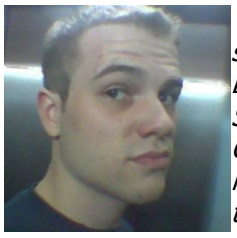
As a product the CCT offers the next functionalities:

- Linguistic corpus management, not only documents as text but also author information and styled document rendering.
- Treatment and validation of TEI encoded documents with support for non-standard characters. It supplies information about the format errors in order to allow the correction by the linguists.
- Intra-documental and collection basic search by single terms.
- Concordance generation (key-word in context) of all the term appearances and location in the document.
- Prefix, suffix and regular expressions search, which is very useful for the linguistic work.
- Phrase search with term distance specification in order to search for linguistic structures.
- Generation of types and tokens lists in document and collection level to allow statistical study of the terms occurrences.

Summarising, the CCT is designed to be scalable and adaptable to the new needs of the corpus compilation process. It is currently an option to manage any TEI encoded corpus and offers the features more often demanded by linguists.

So, here we are

Although this paper was focused in two of our products: NowOnWeb and the Coruña Corpus Tool, we have to mention that now the IRLab is a consolidated group that maintains open several other research lines and also has collaborations with other IR research groups.



Javier Parapar is a Ph.D. student in the IRLab, Department of Computer Science, University of A Coruña. He holds a MS.c. in Computer Science from the same University. His research interests comprise: web IR, news retrieval, recognition and extraction, redundancy detection, summarisation, news clustering, query logging and text fingerprinting.

Álvaro Barreiro holds a Ph.D. from University of Santiago de Compostela. He is an associated professor in the University of A Coruña where he leads the Information Retrieval Lab. He has been the main researcher of several IR research projects funded by the Spanish Government.

Applying IR in a Contextualized Warehouse[4]

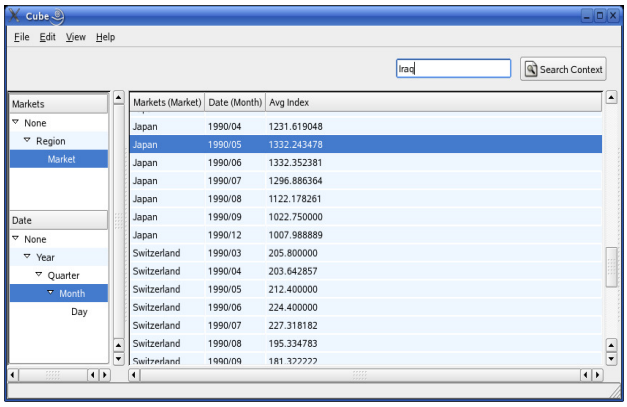
by Juan Manuel Pérez, Rafael Berlanga and María José Aramburu

Current data warehouse and OLAP technologies are applied to analyze the structured data that companies store in databases. The context that help to understand these data are usually described separately in text-rich documents. A contextualized warehouse is a new kind of decision support system where OLAP and IR techniques are combined to integrate a traditional data warehouse and a document repository.

Data Warehouses and OLAP

A data warehouse system stores historical data integrated and prepared for being analyzed by On-Line Analytical Processing (OLAP) tools. Many companies satisfy their needs of strategic information by applying these technologies to their *structured* databases.

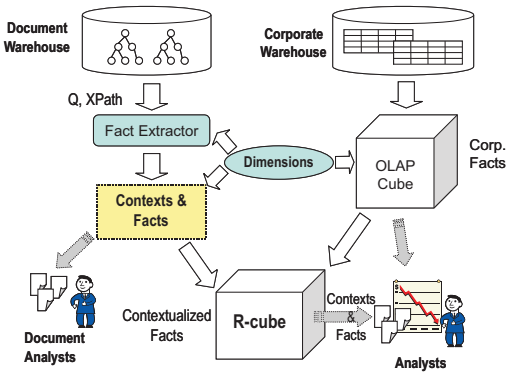
OLAP tools conceptually model information as multidimensional cubes. In this cubes the data is divided into *facts*, the central entities/events for the desired analysis, e.g., the World stock indices, and hierarchical *dimensions*, which provide contextual information for the facts, e.g., the markets (U.K. market, Japanese market, etc.) and the grouping of markets into regions (European region, Southeast Asian region, etc.). Typically, the facts have associated numerical measures (e.g., average stock index) and queries aggregate fact measure values up to a certain level, e.g., average index by region and month, followed by either *roll-up* (further aggregation, e.g., to year), or *drill-down* (getting more detail, e.g., looking at the average per market) operations.



Example traditional data cube for analyzing the stock indices of the major World markets.

The Contextualized Warehouse

A *contextualized warehouse* is a new type of decision support system that allows users to combine all their sources of structured data and documents, and to analyze the integrated data under different contexts.



Architecture of the contextualized warehouse.

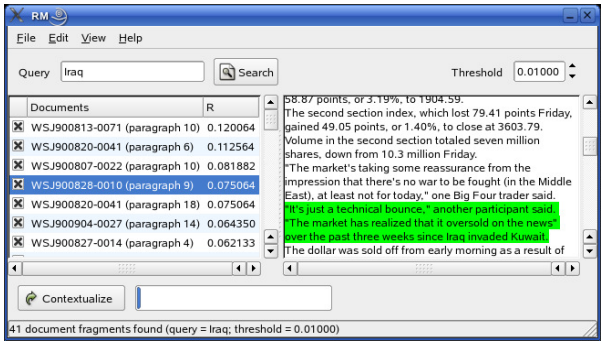
The main components of the contextualized warehouse architecture are a traditional data warehouse, an XML document warehouse and the fact extractor module. The traditional data warehouse integrates the structured data sources (e.g., different stock market databases). The documents coming from external and internal sources are stored in the document warehouse as XML documents (e.g., a collection of business journals gathered from the Web). These documents describe the context (i.e., circumstances) of the data warehouse facts. The document warehouse allows the user to evaluate queries that involve IR conditions. The fact extractor module relates the facts of the traditional warehouse with the documents that describe their contexts. This module identifies dimension

[4] This paper summarizes the PhD thesis of Juan Manuel Pérez, presented in February 2007 and supervised by Dr. Rafael Berlanga and Dra. María José Aramburu.
The authors would like to thank Dr. Torben Bach Pedersen from Aalborg University (Demark). Part of the work presented here was completed with his collaboration.

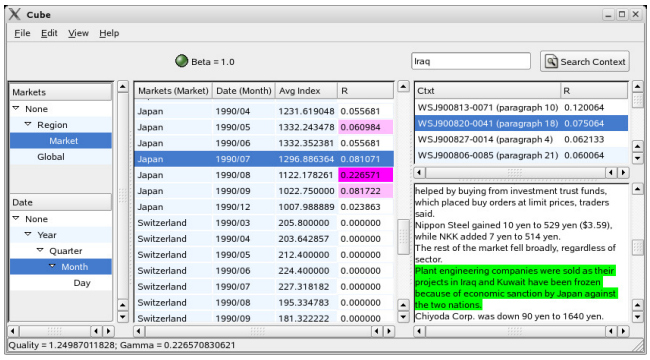
values in the textual contents of the documents and relates each document with the facts that are characterized by these dimension values.

In a contextualized warehouse, the user specifies an *analysis context* by supplying a sequence of keywords (i.e., an IR condition like "Middle East war"). The analysis is performed on a new type of OLAP cube, called *R-cube*, which is materialized by retrieving the documents and facts related to the selected *analysis context*. *R-cubes* have two special dimensions, the *relevance* and the *context* dimensions. Thus, each fact in the *R-cube* will have a numerical value representing its relevance with respect to the specified context (e.g., how important the fact is for the "Middle East war"), thereby the name *R-cube* (Relevance cube). Moreover, each fact will be linked to the set of documents that describe its context.

The relevance and context dimensions provide information about facts that can be very useful for analysis tasks. The relevance dimension can be used to explore the most relevant portions of an *R-cube*. For example, it can be used to identify the markets that were more influenced by the war. The usefulness of the context dimension is twofold. First, it can be used to restrict the analysis to the facts described in a given subset of documents (e.g., the most relevant documents). Second, the user will be able to gain insight into the circumstances of a fact by retrieving its related documents.



Specifying the *analysis context* "Iraq". The retrieved documents are shown on the right.

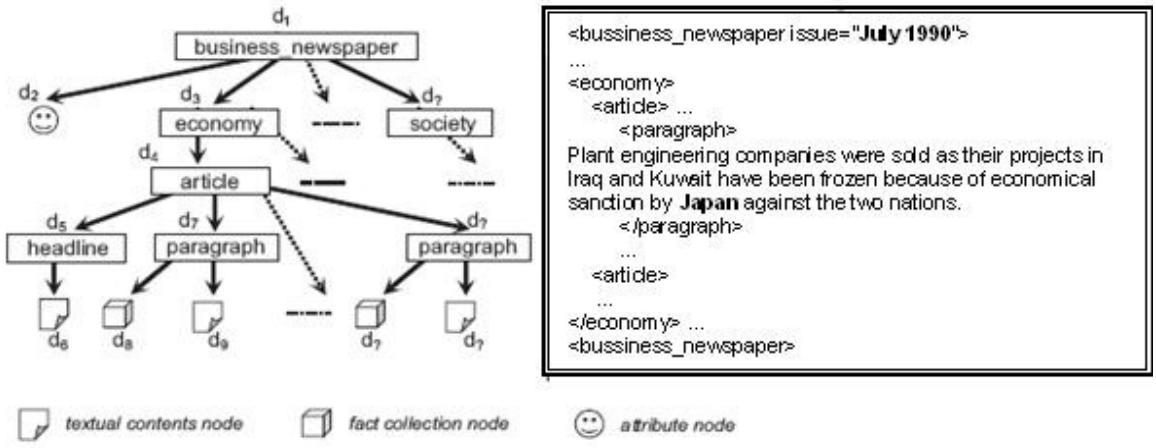


The R-cube presents the facts of the original data cube ranked by relevance (R column), along with the retrieved documents that mention their dimension values (Ctxt).

An IR Model for the R-cubes

Each XML document of the warehouse is represented as tree of nodes. We denote by d_j a node of the tree (d_j stands for document node). There are four different types of document nodes: *textual contents*, *fact collection*, *element* and *attribute* nodes.

- The *textual contents* nodes represent the text sections of the XML documents.
- The *fact collection* nodes are used for representing the facts described in the documents text sections. We model a fact f_i as an n-tuple $f_i = (e_1, e_2, \ldots e_n)$, where each e_j is a value that belongs to the dimension D_j (defined in the warehouse data cubes). For example, $f_1 = (Japan, 1990/07)$ could be attached to a fact collection node that represents the facts described in a text section where the word "Japan" appears. The document that contains this piece of text was published during the month of July 1990.
- Each *element* node depicts the piece of document enclosed within a pair of matching tags of the original XML document. Then, element nodes may have as children other document nodes. These child document nodes can either be attribute, textual contents, fact collection or even element nodes.
- The *attribute* nodes represent the attributes that appear in the XML tags.



Tree-like representation of an XML document.

Ranking the document nodes

In order to rank the document nodes with respect to an IR query, we follow a language modeling approach. The query $Q = q_1 q_2 \ldots q_n$ is considered a sequence of independent keywords q_i, and each document node d_j as a language model. One can see a language model as a black box from which we can sample words. The document nodes d_j are ranked according to the probability $P(Q|\ d_j)$ of obtaining the query keywords when randomly sampling from the respective language model:

$$P(Q \mid d_j) = \prod_{q_i \in Q} P(q_i \mid d_j)$$

$$P(q_i \mid d_j) = (1-\lambda)\frac{TFreq(q_i, d_j)}{\mid d \mid_t} + (\lambda)\frac{ctf_{q_i}}{coll\_size_t}$$

$P(q_i|\ d_j)$ is the probability of sampling the query keyword q_i from the language model of the node d_j. This probability is calculated by smoothing the relative frequency of the query keyword in the piece of text included within the document node. The objective of this approach is to avoid probabilities equal to zero in $P(Q \mid d_j)$ when a document node does not contain all the query keywords. We make the assumption that finding a keyword in a document node might be at least as probable as observing it in the entire collection of documents. Thus, $TFreq(q_i, d_j)$ returns the frequency of the keyword q_i in the piece of text within the document node d_j. $\mid d_j \mid_t$ denotes the total number of words in the text included within d_j. ctf_{qi} is the number of times that the query keyword q_i occurs in all the documents of the collection, and $coll\_size_t$ the total number of words in the collection. The λ factor is the smoothing parameter, and its value is determined empirically, $\lambda \in [0, 1]$.

Ranking the facts

The ranking of facts in the *R-cubes* is performed by adapting relevance-based language models techniques. We estimate the relevance of a fact f_i by calculating the probability $P(f_i|RQ)$ of observing this fact in the set of document nodes RQ Relevant to the query Q:

$$P(f_i \mid RQ) = \frac{\sum_{d_j \in RQ} P(f_i \mid d_j) P(Q \mid d_j)}{\sum_{d_j \in RQ} P(Q \mid d_j)}$$

$P(Q|\ d_j)$ is the probability of observing the query keywords in the document node d_j, which is calculated as discussed above. $P(f_i|\ d_j)$ is the probability of finding the fact f_i in the document node d_j. It is estimated as follows:

$$P(f_i \mid d_j) = \frac{FFreq(f_i, d_j)}{\mid d_j \mid_f}$$

where $FFreq(f_i, d_j)$ returns the number of times that the dimension values of the fact f_i are found in the text sections enclosed within the document node d_j, and $\mid d_j \mid_f$ denotes the total number of dimension values mentioned in d_j.

The set RQ of document nodes relevant to the query Q is built by considering those document nodes that contain at least m query keywords.

We rank these document nodes according to $P(Q|\ d_j)$, and only include in RQ the k document nodes at the top of the ranking.

The Temporal Knowledge Bases Group

The authors of the paper work in the Temporal Knowledge Bases Group (TKBG) of the Jaume I University, Castellón (Spain). The team is currently composed by eight members from Jaume I University and three external researches.

Some members of TKBG, from left to right: Roxana, Lola, María José, Juanma, Rafa, Isma, Victoria and Mari Paz.

The TKBG is mainly concerned with the investigation of novel methods for modelling, querying and managing very large collections of semi-structured documents. Our approach for the effective exploitation of these document collections combines techniques stemming from several research areas: Databases, Information Retrieval, Natural Language Processing and Pattern Recognition.

The main active research lines started by the group around this problem are the following ones:

- Document management, storage and retrieval.
- Knowledge(-based) extraction.
- Ontology Learning.
- Multidimensional analysis of semi-structured documents.
- Semantic GRID.

More information about the TKBG can found at http://krono.act.uji.es.

Juan Manuel Pérez *obtained the B.S. degree in Computer Science in 2000, and the Ph.D. degree in 2007, both from Universitat Jaume I, Spain. Currently he is associate lecturer at this university. Contact him at JuanMa.Perez@lsi.uji.es.*

Rafael Berlanga *is an associate professor of Computer Science at Universitat Jaume I, Spain. He received the B.S. degree from Universidad de Valencia in Physics, and the Ph.D. degree in Computer Science in 1996 from the same university. Contact him at berlanga@lsi.uji.es.*

María José Aramburu *is an associate professor of Computer Science at Universitat Jaume I, Spain. She obtained the B.S degree from Universidad Politécnica de Valencia in Computer Science in 1991, and a Ph.D. from the School of Computer Science of the University of Birmingham (UK) in 1998. Contact her at aramburu@icc.uji.es.*

Search Oriented Transform of Web Sites

By César Llamas, Pablo de la Fuente and Jesús Vegas

Introduction

The great amount of information supported nowadays in Internet, has increased the relevance of web sites searchers. These sites serve as facilitators of first access web pages, also web sites, to start navigate in order to find some resource. In practice, usual structure of web sites is organized with the page concept in its core. Page information is displayed in textual or graphical manner, along with several interaction elements, in order to maximize business opportunity with a potential standard client if the aim is interaction with a service oriented enterprise. Otherwise, in a more general case, to enhance information accessibility would be the goal. However, as a consequence, web crawlers encounter severe problems indexing many of these sites because their interaction appearance, and produce bad query search results.

It is known that about 88 % of the times that an average user starts using the web he initiates the task querying a search site. In other hand, one of the ten basic rules for a good practice of web design is to avoid too large pages, splitting them whenever it its possible forming a hierarchy of pages related by hyperlinks. However, this partitioning on pages has some other negative effects, being one of them derived from that web search engines organize searches in a page based data model. Its notion of document is limited to the physical limits of one page. In summary, user experience and personal abilities are conditioning the success of the search.

Taking this into account, algorithms and procedures are needed not only to provide users with the best approach to match queries but also to web search engines, in order to a better navigation and indexing. This implies that a good solution to start navigation from a query must take into account the structural relations and the hierarchy presented in hyperlinks.
Our proposal asserts that web search engines must be provided a modified view of the original web sites, adapted to the way in what this services crawl, analyze and index the information hold in web pages and hyperlinks.

Exposition and Methodology

Under our proposal lies the assumption that it is possible to render web pages to web crawlers in a way adapted to its internal procedures of indexing and searching, usually very different to the final user view. Therefore, we pose to exploit the internal web site link structure offering a better approach to multiple word queries, by the means of an *ordered conjunction of interlinked pages*. This hypothesis would be discriminated against two different well spread out kinds of sites, whose results could be extrapolated to some other structures.

There exist several ways in what query results could be enhanced taking advance of some features of the user query. Here we deal with multiple word queries, and with the fact that these terms could be scattered out in several linked pages that in fact make up a one document far from the limited view of one ordinary page.

Reorganization in NWEB

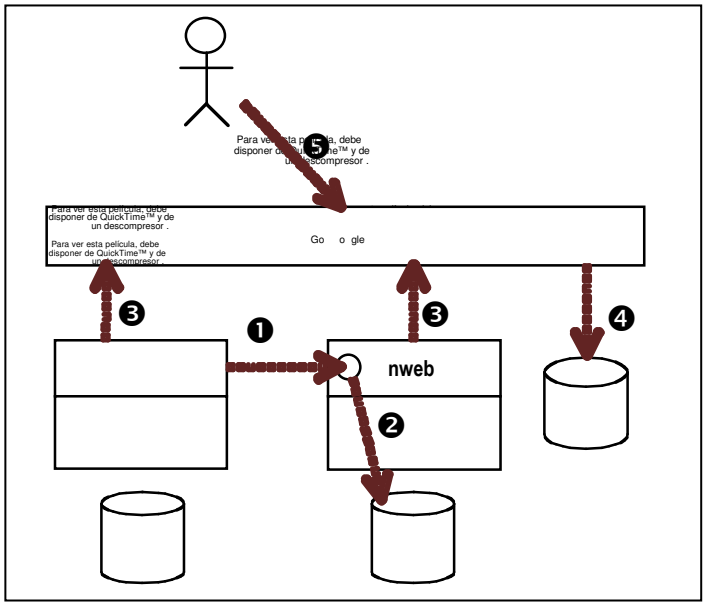
We call *nweb* the rendering obtained from the original web site, from a process of reorganization in what new document units are formed by groups of linked physical pages up to a certain degree *n*. Roughly speaking, this alternative view is what will be given to the web crawler of the search site.

In this way, we could speak about an nweb of certain page a as a new page compound by itself and the set union of the contents of every paged linked from it.
The concept of nweb could be recursively applied started from a given document, to obtain increasingly general documents starting from an original document (web page) that we consider as a seed.

The Relevance Hypothesis

The main hypothesis behind all this work is that when the information is contained in two or more web pages linked among them, the relevance of the nweb that includes these pages is greater than the sum of the relevance of the each web page that compounds it.

To validate this hypothesis we have build a system for experiment with the nweb concept applied to several real web sites. This platform allows creating alternative web sites from a web site, organizing the information in nweb controlling the depth of them. Eventually, these sites, the original and its nweb versions (nweb2, nweb3, etc.; depending on the depth) are indexed by Google, which plays the role of judge assigning the relevance and answering the questions over the sites under study, allowing establishing comparisons between the results obtained by the different nweb sites and the original one. Three cases can be observed: (1) the original site appears in the results list higher than it's nweb versions, which indicates that the original information distribution is better; (2) the original site appears behind some of it's nweb version in the results list, making the nweb version more suitable than the original one; and (3) the results list returned by the search engine does not include any of the sites under study, situation that it is not relevant to our purposes. The next figure represents the scheme of the experimental system.



In this experimental phase two different web sites have been chosen, which could be a good representative of some others: the documentation site of the package "java.util.jar" of the programming language Java, and one wiki-book. The first site is highly connected with many hyperlinks among the pages. The second site shows a highly hierarchical structure.

Experimental Results

The original sites and the "nweb" versions (depth from 2 to 5) were exposed to the Google indexing system; after that, we build some queries to the Google search system. Queries were formed by simple as well as complex terms, all of them extracted from the original web sites and combined to obtain queries with length from 2 to 5 terms. To explain details about the query process exceed the purpose of this document.

The first column of Table 1 shows how many times each site under study appears in the result list returned by the web search engine before others. The second column shows how many times our pages were not ranked by Google. 500 queries for the five sites in study were made. Although the two sites selected in this experiment javadoc and wikibook were studied independently each other, results were combined here to obtain wider conclusions.

As can be seen from this table, nweb is the best organization of the content of the web site than the original site about 73.6 % of the cases; original site wins only in 10.8 % of queries. In the nweb set, nweb2 wins in the 50.8 % of the cases, outstanding on all other nweb situations.

In relation to the two kind of sites under study, it can be said that our nweb proposal works better with a lower degree of connectivity (wikibook alike), although with a javadoc site alike works also well.

Conclusions and Future Work

In this paper we study the convenience to render distinct views of the web sites, in relation with the user intention: one browser oriented and other one to index and search oriented. In this work we have show how can be reorganized a web site to present a more appropriated view to the web search engines, especially when the query terms are scattered over several web pages bounded by hyperlinks. Also the notion of nweb has been defined, along with an experimental environment.

	Original	Not found	Total Nweb	Nweb2	Nweb3	Nweb4	Nweb5
2 terms	27	19	78	**59**	7	10	2
3 terms	9	16	83	**60**	9	9	5
4 terms	11	17	100	**62**	11	15	12
5 terms	7	26	107	**73**	6	17	11
Total	54	78	368	254	33	51	30

Table 1. Experimental results for the original site and its nweb views. The third column shows how many times some nweb appears before the original site in the result list.

It could be said that, from all the nweb views, the nweb2 presents the best results in the experiments, and it contributes tracks about how can be done the searcher view. More work has to be done related to the complete definition of the search view.

In order to integrate the view concept in the web information retrieval process a lot of work has to be done, to obtain user transparency. In this work we have focused in the significance of our proposal, but a further study about the implications of this application in the interactive process of the web search is needed.

Forthcoming Events

Edited By Andy MacFarlane

30th European Conference on Information Retrieval (ECIR 2008)
The annual Conference of the IRSG, Glasgow, UK.
30th March - 3rd April, 2008.
http://ecir2008.dcs.gla.ac.uk/

ACM Conference on Electronic Commerce (EC'08)
Of interest to members working in the area of Spam control, web search etc.
Chicago, Illinois, 8th-12th July, 2008
http://www.acm.org/sigs/sigecom/ec08

AAAI 2008 Workshop - WIKIPEDIA AND ARTIFICIAL INTELLIGENCE: AN EVOLVING SYNERGY
Wikipedia as a source, with various themes of interest such as semantic web and cross language.
Chicago, Illinois, 13th-14th July, 2008
http://lit.csci.unt.edu/~wikiai08

Second International Workshop on Scalable Data Management Applications and Systems (SDMAS'08) to be held within The 2008 International Conference on Parallel and Distributed Processing Techniques and Applications
A conference focused on scalability issues, of interest to members working on large scale IR problems such as web search.
Las Vegas Nevada, USA, 14th-17th July, 2008
http://www.arcos.inf.uc3m.es/~jdaniel/sdmas08/

8th Industrial Conference on Data Mining (ICDM 2008)
Of interest to members working in the area of text mining.
Leipzig, Germany, 16th-18th July, 2008.
http://www.data-mining-forum.de/

The 31st Annual International ACM SIGIR Conference (SIGIR 2008)
The big annual IR get together for researchers all over the world.
Singapore, 20th-24th July 2008.
http://www.sigir2008.org/

The Seventh International Conference on Mathematical Knowledge Management (MKM 2008)
A workshop with various themes of interest including digital libraries and search/retrieval for mathematical knowledge.
Birmingham, UK, 28th-30th July 2008
http://events.cs.bham.ac.uk/cicm08/mkm08/

5th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH2008)
Of interest to members working in the area of areas such as recommender systems
Hannover, Germany, 28th July - August 1st 2008.
http://www.ah2008.org/

Fifth International Conference on Visual Information Engineering (VIE'08)
The more visual aspects of IR is an important theme of this conference.
Xi'an, China, 29th July - 1st August 2008.
http://vie08.qmul.net/

The 10th International Conference on Music Perception and Cognition (ICMPC10)
Of interest to members interested in the cognitive aspect of music retrieval.
Hokkaido University, Sapporo, Japan, 25th-29th August 2008
http://icmpc10.psych.let.hokudai.ac.jp/

DEXA 2008
A collection of various conferences with themes on IR.
Turin, Italy, 1st–5th September 2008.
http://www.dexa.org

Ninth International Conference on Music Information Retrieval (ISMIR 2008)
Of interest to members who work in the area of music retrieval
Philadelphia, USA, 14th-18th September 2008.
http://ismir2008.ismir.net/

2008 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC'08)
A general visualisation conference of interest to members working in visualisation and search.
Herrsching am Ammersee, Germany, 16th-20th September 2008
http://vlhcc08.cs.unibw.de/

Information Seeking in Context 2008 (ISIC 2008)
Of interest to members who work in the area of context and IR
Vilnius, Lithuania, 17th-20th September 2008
http://www.kf.vu.lt/isic2008/

Second Information Interaction in Context Symposium (IIiX 2008)
Of interest to members who work in the area of context and IR
BCS Covent Garden, London, 14th-17th October 2008.
http://irsg.bcs.org/iiix2008/index.php

ACM Seventeenth Conference on Information and Knowledge Management (CIKM 2008)
A conference with a number of major themes of interest to members including IR and information management.
Napa Valley Marriott Hotel & Spa, California, October 26th-30th 2008.
http://www.cikm2008.org/

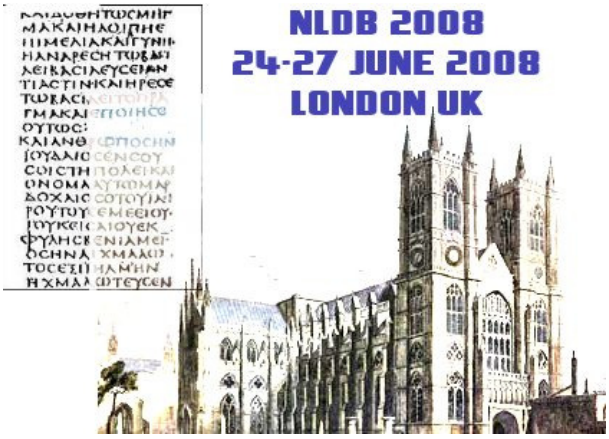
Contacts

Web: http://irsg.bcs.org/
Email: irsg@bcs.org.uk
Subscriptions: http://irsg.bcs.org/membership.php
ISSN: 0950-4974

To subscribe, unsubscribe, change email address or contact details please visit http://irsg.bcs.org/ or email irsgmembership@bcs.org.uk.

The IRSG is a specialist group of the British Computer Society.
To automatically receive your own copy of Informer, simply join the IRSG via the IRSG website.

13th International Conference on Applications of Natural Languages to Information Systems
http://www.nldb.org



Natural language has always posed serious challenges to theory of computation and Turing Machine based intelligence. This controversy also inspired the writing of exciting stories in the literature, e.g., *The Cambridge Quintet* by John Casti. However, natural language plays a key role in many areas such as

Web and *Web based Information Systems*
Information Retrieval in Cross-Lingual or Mono-Lingual settings
Digital Libraries
Documentation in Software Engineering
Document and *Content Management Systems*
Human-Machine Interaction
Machine Translation
Knowledge Management and *Electronic Encyclopaedias*

Since 1995, the NLDB conference, with high impact factor, has aimed at bringing together researchers, industrials and potential users interested in various applications of Natural Language in the Web and database driven information systems area. It has contributed to many areas such as

- improving the development process from the viewpoint of developers (e.g., the process of requirements engineering, conceptual modelling, validation, etc.)
- usability of applications (e.g., natural language query interfaces, retrieval, semantic web, etc.)
- Knowledge extraction and dissemination (e.g., text mining, knowledge discovery, etc.)

To highlight these inspiring connections, NLDB 2008 will take place from June 24 to June 27 in the world city of London (UK). It will provide the stage for fruitful and challenging discussions.

Indicative Topics of Interest

- **Semantic Web and Information Retrieval**
- **Text and Web Mining**
- **Taxonomies and Ontology Extraction from Text**
- **Document Classification and Indexing**
- **Natural Language in Conceptual Modelling**
- **Natural Language in Software Engineering**
- **Natural Language Based Interfaces for Database Querying and Retrieval**
- **Natural Language Based Integration of Systems**
- **Large Scale On-line Linguistic Resources, Electronic Dictionaries, Digital Libraries**
- **Applications of Computational Linguistics in Information Systems**
- **Management of Textual Databases**
- **Natural Language for Data Warehouses and Data Mining**

University of Westminster