Informer

Number 23, Summer 2007
ISSN 0950-4974

BCS INFORMATION RETRIEVAL

In This Issue

Editorial 1
by Tony Rose

Forthcoming Events 2
Edited by Andy MacFarlane

Research Updates

"Searching for the Music You Like" 4
by Maria M. Ruxanda

"Exploring Maps through Georeferenced Images and RDF Shared Metadata" 6
by Carlo Torniai

"Using Semantic Relations to improve Question Answering" 8
by Isabel Segura-Bedmar

"Modeling & Annotation of Dance Media Semantics" *by Kannan Rajkumar* 11

Book Review 12
"Multimedia Data Mining and Knowledge Discovery"
Reviewed by Andrew Neill

Obituary: Karen Spärck Jones 14
by Steve Robertson

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.

Let me make an admission. In private, I've sometimes held a slightly condescending view of academic research, believing (quite unfairly, of course) that the relevance of some of the more "blue skies" work to the lives of most people working in the search industry was peripheral at best. So when Leif Azzopardi (our Chair) approached me suggesting we run a special issue devoted almost exclusively to academic research, I had to confess to a slight degree of unease.

But how wrong could I be. In this issue, we present a selection of Research Updates submitted by attendees at the Summer School on Multimedia Semantics. Academic they may be, but these articles are also undeniably engaging, topical and thought provoking. What's more, the connection between these emerging research ideas and many of the current "Web 2.0" success stories is quite evident. None more so the article on music search – as you may know, last.fm began life just a few years ago as a student project at Southampton University, but was recently acquired by CBS for the princely sum of $280m. Now that's what I call a return on investment.

On a more reflective note, we also bring you Steve Robertson's tribute to Karen Sparck Jones, who died earlier this year. She was a true pioneer in our community, and her loss is made all the more poignant coming so soon after being awarded the 2007 Lovelace Medal.

We should also show our appreciation to Andrew Neill, who contributes yet another excellent book review, this time of "*Multimedia Data Mining and Knowledge Discovery*", by Petrushin and Khan. By the way, we also have a copy of Martin White's "*Making Search Work*" available for review – if you're interested, just drop us a line at irsg@bcs.org.

And lastly - were you at Search Solutions 2007? If so, you'll know what a success it was, attracting twice as many delegates as last year and filling the venue to capacity. If you missed it, fear not: just visit the IRSG website, where you can find all the talks, photos, podcasts and more. Until next time,

All the best,
Tony

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

Forthcoming Events

Edited By Andy MacFarlane

4th International Workshop on Text-Based Information Retrieval" (TIR-07)
A Workshop in conjunction with the 18th International Conference on Database and Expert Systems Applications, DEXA 2007, Regensburg, Germany, 3-7 September 2007.
http://www.aisearch.de/tir-07/

International Conference on Future of Knowledge Organisation in the Networked Environment (IKONE 2007)
Bangalore, India, 3-5 September 2007. A conference focusing on the issue of knowledge organisation for IR.
http://www.drtc.isibang.ac.in/ikone/

7th International Conference on Knowledge Management (I-Know'07)
Graz, Austria, 5-7 September 2007. A generic Knowledge Management conference.
http://i-know.know-center.tugraz.at/

1st International Conference on New Media Technology (I-Media'07)
Graz, Austria, 5-7 September 2007. A conference of interest to members who work in the area of Web 2.0. http://www.i-media.tugraz.at/

Eighteenth International ACM Conference on Hypertext and Hypermedia (HT 07)
Manchester, U.K. 10-12 September 2007. A conference of interest to members who working the area of Hypertext and IR.
http://www.sigweb.org/ht07/

10th International Conference on Digital Audio Effects (DAFx'07)
Bordeaux, France, 10-14 September 2007. A music digital audio conference with a theme on music IR.
http://dafx.labri.fr/

11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2007)
Budapest Hungary, 16-21 September 2007. The main European Digital Library conference, together with the CLEF workshop. www.ecdl2007.org

Ubicomp 2007
Innsbruck, Austria, 16-19 September 2007. A Ubiquitous computing conference of interest to members who focus on mobile search.
http://www.ubicomp2007.org/

First IEEE International Conference on Semantic Computing (ICSC 2007)
Irvine, California, U.S.A, 17-19 September 2007. A generic semantic computing of those members with a focus on issues such as the semantic web, NLP etc. http://icsc2007.eecs.uci.edu/

1st Workshop on Social Information Retrieval for Technology-Enhanced Learning (SIRTEL'07)
Crete, Greece, 17-20 September 2007. A conference on using social networking techniques (e.g. democratic indexing) in order to support information retrieval. http://ariadne.cs.kuleuven.be/sirtel

Location and Context Awareness (LoCA 2007)
Oberphaffenhofen, nr Munich, Germany, 20-21 September 2007. A symposium associated with Ubicomp 2007, and which will of interest to those who work in mobile search.
http://loca2007.context-aware.org/

11th International Symposium on Database Programming Languages (DBPL 2007)
Vienna, Austria, 23-24 September 2007. A database focused symposium with several themes on IR.
http://dcc.puc.cl/dbpl07/

8th International Conference on Music Information Retrieval (ISMIR 2007)
Vienna, Austria, 23-27 September 2007. An increasingly popular conference on music IR.
http://ismir2007.ismir.net/

Workshop Information Retrieval 2007
University of Halle, Germany, 24-26 September 2007. A general workshop in IR, held by our professional colleagues in Germany.
http://lwa07.informatik.uni-halle.de/fgir07/fgir07_en.htm

Current Trends in Computer Science: Annual International Conference of the Mexican Computer Science Society (ENC 2007)
Morelia, Mexico, 26-28 September 2007. A Mexican Computer Science conference with a track on IR.
http://enc.smcc.org.mx/english/index.php?option=com_content&task=view&id=17&Itemid=28

Towards Genre-Enabled Search Engines: The Impact of NLP
Borovets, Bulgaria, 30 September 2007. A workshop of interest to members interested in Natural Language Techniques for IR.
http://www.sics.se/use/genre-ws/ (conference web site: http://lml.bas.bg/ranlp2007/)

IADIS INTERNATIONAL CONFERENCE WWW/INTERNET 2007
Vila Real, Portugal, 5-8 October 2007. A web conference of interest to members working in the area of Web Search.
http://www.internet-conf.org/

2nd International Conference on Metadata and Semantics Research (MTSR 2007)
Ionian Academy of Corfu (Ionian University), Corfu, Greece, 11-12 October 2007. A meta-data conference of interest to members who apply knowledge organisation techniques to IR.
http://www.mtsr.ionio.gr/

9th Russian Conference on Digital Libraries
Pereslavl, Russia, 15-18 October 2007. A conference of interest to members working in the area of Digital Libraries.
http://rcdl2007.pereslavl.ru/en/news.shtml

International Conference on the Theory of Information Retrieval (ICTIR 2007)
Budapest, Hungary, 18-20 October 2007. A conference focusing on the theoretical side of IR.
http://www.infota.org/ictir2007/index.html

ACM Recommender Systems conference (RecSys 2007)
Minneapolis, Minnesota, USA, 19-20 October 2007. A conference of interest to members who work in collaborative IR. http://recsys.acm.org/

14th String Processing and Information Retrieval Symposium (SPIRE 2007)
Santiago, Chile, 29-31 October 2007. A major string processing and information retrieval conference.
http://www.cwr.cl/spire2007/

5th Latin American Web Congress (LA-Web 2007)
Santiago, Chile, 31 October – 2 November 2007. A south American web conference associated with SPIRE 2007. http://www.cwr.cl/la-web2007/

2007 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'07)
Silicon Valley, USA., 2-5 November 2007. A multifaceted conference with many themes of interest to IRSG members.
http://www.cs.sjsu.edu/wi07/

ACM Sixteenth Conference on Information and Knowledge Management (CIKM 2007)
Lisbon, Portugal, 6-10 November 2007. A big Knowledge Management, Information Retrieval and Database conference.
http://www.fc.ul.pt/cikm2007/

15th ACM International Symposium on Advances in Geographic Information Systems (ACM GIS 2007)
Seattle, Washington, USA. 7-9 November 2007. A GIS conference of interest to members working in the field of Geographical Information Retrieval.
http://www.cise.ufl.edu/dept/acmgis2007/

Workshop on Text Mining and Applications (TeMA-2007)
Guimares, Portugal, 3-7 December 2007. A workshop of interest to members working in areas such as Information Extraction and Question/Answering.
http://epia2007.appia.pt/index.php?option=com_content&task=view&id=54&Itemid=74

The 6th International Semantic Web Conference (ISWC 2007)
Busan, South Korea, 11-15 November 2007. A combined Semantic web and Digital Library conference, of interest to members who work in IR for either of these fields.
http://iswc2007.semanticweb.org/

International Workshop on Collaborative Knowledge Management for Web Information Systems (WE.KNOW 2007)
Nancy, France, 3 December 2007. A workshop of interest to members interested in social networking methods to support IR.
http://isweb.uni-koblenz.de/weknow2007

The 10th International Conference on Asian Digital Libraries
LOCATION:Hanoi, Vietnam, 10-13 December 2007. A conference of interest to Digital Library researchers and practitioners.
http://icadl2007.vista.gov.vn/

Research Update:

Searching for the Music You Like – Current Trends

By Maria M. Ruxanda

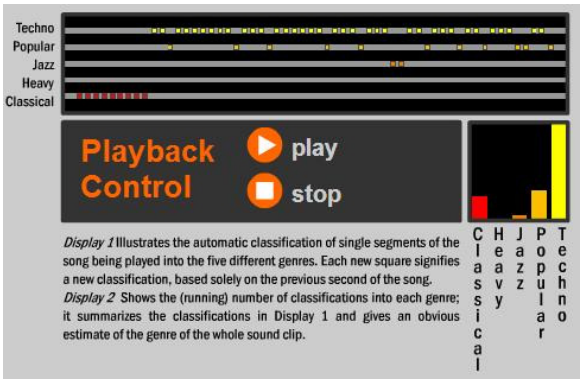
The intelligent Internet has already revolutionized information availability for the private user. Radio and TV broadcasting are now in the digital age and the big record companies are selling music on-line on the web. Digital music is becoming more and more widespread as personal collections of music grow to thousands of songs. Consequently, music search engines play an important role in entertainment services and music sales.

The most popular are and will be the music search engines that search for "similar music" striving to provide natural and intuitive music retrieval based on the media content itself. Along with the already known Last.fm (www.last.fm), Pandora (www.pandora.com) and MusicIP (http://musicip.com), the on-going Intelligent Sound Project (www.intelligentsound.org) is part of this trend.

Music Search Engines Providing Similarity-based Search

The concept of music similarity has been generally dealt with from two perspectives. The first includes techniques that use collaborative filtering e.g. as done by Last.fm. Here, past user behavior is used for guidance. For example, each time a user listens to a piece of music on the streamed radio stations or on his own computer or iPod, this is recorded. Last.fm builds a detailed profile of each user's musical taste and then it becomes possible to offer guidance akin to "those who liked this music also like this other music". Common ways of recording collaborative data is in the form of playlist co-occurrence relative to songs or artists that often appear in the same playlist, or recording how users are rating different songs. The advantage of collaborative filtering is that it scales well to large music databases. However, it does not work well when new music is introduced.

The second perspective includes music content-based techniques e.g. as done by MusicIP. Music content can be represented by the so called "features" which basically can be seen as a numerical representation of the sound. For instance, in MusicIP the extracted features appear in the form of unique acoustic fingerprints. The songs can be identified by these acoustic signatures regardless of the file format or data compression rate and thus it becomes possible to automatically discover tracks that are perceivably similar. The content-based techniques can easily classify the new music and can be integrated with collaborative filtering strategies. Generally, they are automatic and scale well to large music collections. An exception here is the popular search engine Pandora which performs music similarity search based on manually extracted features. A team of fifty musician-analysts listen to every song at a time, studying and collecting hundreds (close to 400) of musical details like melody, harmony, instrumentation, rhythm, vocals, etc. However, the price to pay is the low scalability.



The ISP Clever Jukebox – Automated Genre Classification

The on-going Intelligent Sound Project (ISP) envisions a framework for the next generation of intelligent sound services that will provide both automatic methods for feature extraction, classification and segmentation, and new techniques for efficient similarity search methods that integrate the querying of collaborative data with content-based search. One of its very first tasks, already completed, is the creation of a system for automatic musical genre classification

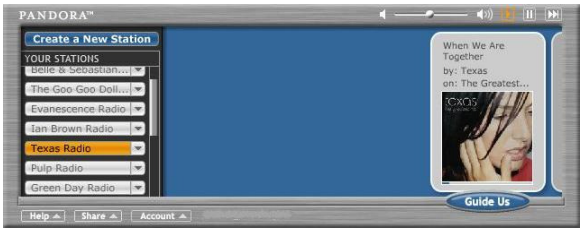
(www.intelligentsound.org/demos/automusic.swf). The genre classification is achieved by using frequency-based extracted sound features followed by a linear regression classifier. The extracted features are in the form of a 30 dimensional feature vector on the time-scale of 1 second, so that a song can be classified into a genre for each second of played music.

Music Search Engines Deliver Playlists to the Users

Regardless of the manner in which the music similarity is dealt with, the popular music search engines are focused upon the concept of playlist. Users are searching for the music they like and they are receiving playlists or mixes of single songs from a variety of artists. The playlist concept is disguised under different forms and names. In Last.fm and Pandora, a playlists is a user-personalized "radio station": one song is presented to the user at a time and the user can skip to the next song, but he can't explicitly choose to listen again to a song already played.

The Radio Station in Last.fm

Nevertheless, the user can continuously give feedback and rate songs (e.g. "I like it", "I don't like it"), so that the radio station is dynamically adapted to his taste. In MusicIP, a playlist is called a "mix" and can be seen as a regular playlist in the sense that the user can see the entire playlist at any time and can listen to the songs in any order he wants. But it can be customized to the user's taste and updated according to his feedback - the user has the options to replace in any playlist the albums, artists or songs he does not like or request more songs similar to his favourite songs or less songs similar to those he dislikes.



The Radio Station in Pandora

Which type of playlist is better is a question with more than one answer. Users are various and subjective. Some prefer the web based radio stations (like in Last.fm, Pandora) where the streaming music keeps coming to their ears and all it takes for further customization of the playlist is pushing a rating button. Other prefer the regular playlists - even if more effort may be required for customization, more control is allowed on what and when should be played from personal collections of music (like in MusicIP). And other like both types of playlists, and switch between them according to mood and usage context.

What we can treat as a certain fact is the increased popularity of such music search engines, and consequently the fact that more and more users are not listening to albums but to playlists containing songs from various artists. It seems that albums are fading away

The Custom Playlist in MusicIP

on the music selling market while the playlists are strongly emerging as the new items desired by the consumers, as recently stated by an article in the New York Times ("The Album, A Commodity in Disfavor", 26 March 2007).

Therefore, part of the ISP's goals and research directions is to formalize a more general framework for music querying by defining all types of playlists and operations on these playlists, consequently aiming towards a similarity algebra for music data. Such an algebra will provide the necessary set of operations to support similarity queries in a large music database while treating playlists as core query units. The on-going Intelligent Sound Project is coming along with the current trends providing intelligent sound services and is striving to go even one step further to meet the emerging demands and complement what is missing.

Maria Magdalena Ruxanda is a PhD Student within the Database Research Group, at Computer Science Department, Aalborg University, Denmark. She is a member of the Center for Data-intensive Systems (http://daisy.aau.dk/) and her research interests are focused on multimedia databases and information retrieval. Her PhD study is funded (in part) by the Danish Research Council for Technology and Production Sciences project No. 26-04-0092 Intelligent Sound (www.intelligentsound.org). She can be contacted at mmr@cs.aau.dk.

Research Update:

Exploring Maps through Georeferenced Images and RDF Shared Metadata

By Carlo Torniai

What if you could browse vicariously pictures, locations, events, based on their locations and relationships simply using a map as multi-layered interface?
Our work explores the kinds of metadata that can be captured at the time a photo is taken, and ways to link photos, locations and users together according to these metadata in order to make it easy for people to browse the "big picture". The objective of our work is to create an experience where a user view a photo on the web, then jump, for instance, to other photos in the field of view or taken nearby, or look for other people that took pictures of that location.

It draws on the network effect of the web by including not only the user's own photos but any photo that can be discovered with suitable metadata. This includes location (GPS or other mobile location), heading information to identify the position and direction of the camera, details about date, time and author. The photos are taken by different people and are shared on the web. The key to this linking is metadata attached to the photo. There are no explicit hyper-links between photos, making it easy for people to contribute. Automatic linking is achieved by the discovery of photos on the *semantic web*.

Distributed Environment for Metadata Sharing

Our work is focused on data that can be captured at the time a photo is taken and on mechanisms to share and relate them in order to provide a novel map navigation experience.

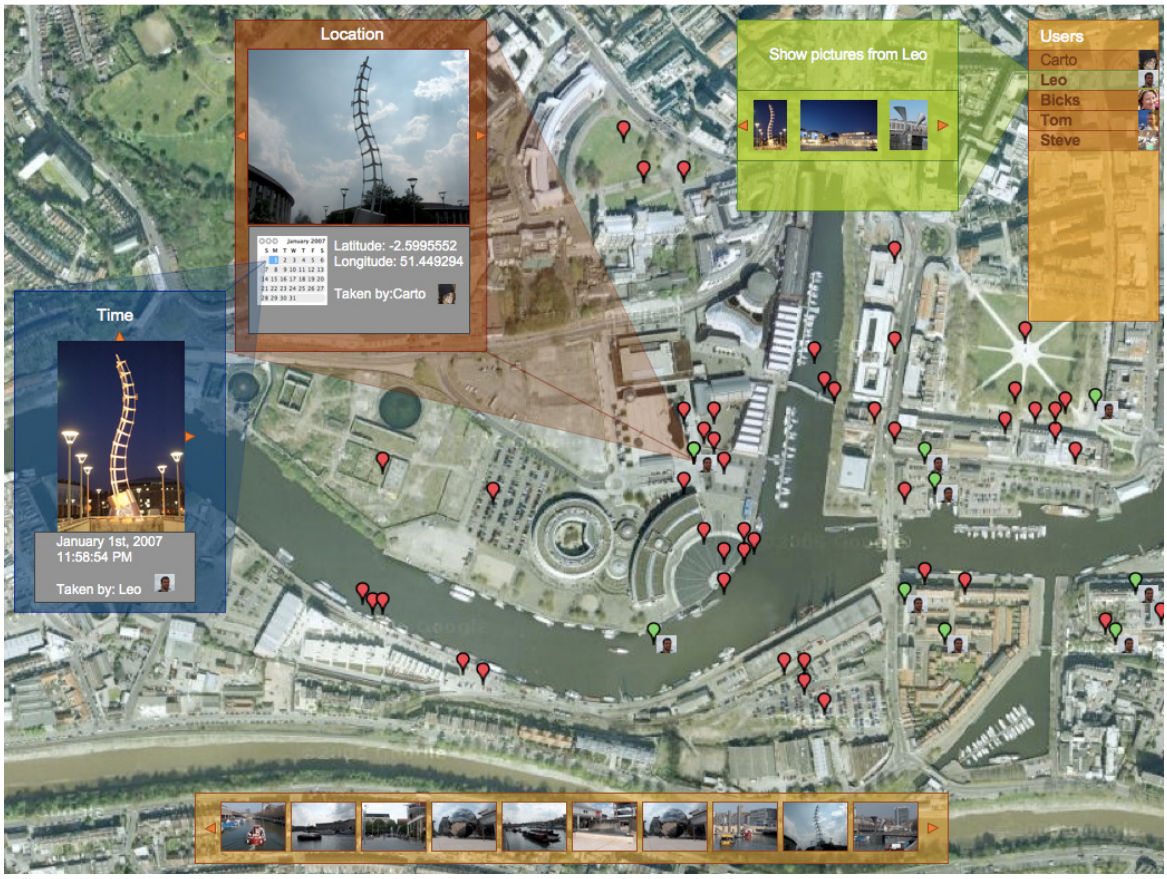


Fig. 1 – An example of the "mashup-interface"

The main idea is to use metadata related to pictures and photo collections and share them in a distributed environment.
Metadata are described by Resource Description Framework, (RDF - http://www.w3.org/RDF/), a metadata standard for representing digital resources. Spatial relations between nearby pictures, together with other relations that can be exploited from metadata, are discovered by means of inference over their RDF descriptions.

The distributed environment for metadata sharing and discovering can be based on SPARQL (http://www.w3.org/TR/rdf-sparql-query/) endpoints that can gather metadata from Flickr (http://flickr.com) photosharing service.

Google Maps (http://maps.google.com) can be then used to gather all the information in a browsable interface. This interface represents a new paradigm of interaction with maps providing different ways of touring and exploring locations, pictures, events and browse information related to users. Once different users can share metadata about their photo collections, and when mechanism to relate pictures, locations, events and users according to these metadata are provided, maps can be transformed in a multi-layered virtual space with different levels of interaction. Pictures can be the entry points to move towards other pictures and locations, look for other places, discover events and meet people: the city map, the users and the georeferenced pictures together with their metadata, become part of one big *"mashup interface"* (Fig 1).

In the picture our view of map interaction is depicted. The exploration interface is created as a mashup between Flickr photosharing services, Google Maps and shared RDF metadata repositories. The map becomes an "interactive desktop" enabling location/time/event/user-based navigations of cities through geotagged pictures. In the example depicted, the user is currently looking around Bristol Millennium Square touring the place through pictures taken from different

users. The map displays available georeferenced pictures in the nearby area. The user wants to see the same place at night, during the beginning of the year eve. Switching seamlessly from location-based to time-based browsing a picture of the place at that particular time and date is shown. The picture has been taken by Leo. The user can enable the user-based navigation looking for other pictures taken nearby by Leo that are displayed in the map.

Conclusions

Based on the distributed architecture described a testing environment for metadata sharing and relations discovering has been implemented so that users photo collections are enhanced by relations with other users pictures. Based on metadata expressed in RDF, it is possible to build a service for discovering, linking and browsing geographical related photo in a novel way.

We have identified the possibility to:
• create *mashups* between photosharing services (Flickr) maps services (GoogleMaps) and shared metadata repositories (SPARQL endpoints) in order to enable multilayered *"mashup interfaces"* for exploring cities;
• use RDF as common format for representing geotagged resources and metadata;
• rely on the *semantic web* as mean for sharing RDF metadata in order to provide automatic relations discovering and to enable a dynamic location/time/event/user/-based navigation of cities.

Dr. Carlo Torniai works at Media Integration and Communication Center, Univesrità di Firenze, Italy. He earned a Master in Internet Engineering in 2003 and a PhD in Computer Science in 2007. He is interested in pattern recognition, automatic video annotation and semantic web application related to multimedia content. He can be contacted via:
torniai@micc.unifi.it

Research Update:

Using Semantic Relations to improve Question Answering

By Isabel Segura-Bedmar

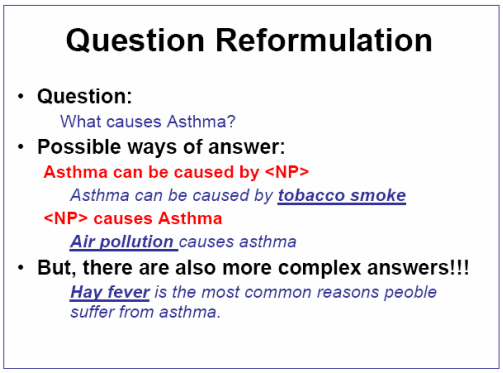
In the last years, World Wide Web is becoming a universal and essential knowledge source accessible to everybody. Billions of queries are submitted daily to the web search engines. A large quantity of data is available and to find concise answers has become a difficult task.

The main goal of Information Retrieval (IR) is to find relevant documents among large collections of data, whereas Question Answering (QA) systems try to find concrete answers for user's natural language questions. QA systems are very useful when user needs to know the exact answer and does not want to review all the documents related to the topic. Besides, QA is particularly useful for less visually rich interfaces such as speech driven interfaces.

"Different types of questions require different strategies to locate the answer"

QA research deals with a wide range of questions types such as factoid (asking for the name of a person, a location, the day on which something happened, etc), definition (What / Who is X?), list (one answer containing a list of items), more complex questions such as How, Why, and so on. Different types of questions require different strategies to locate the answer. Various QA systems use a Question Reformulation module which tries to identify the possible ways of answering a given question. This module often uses patterns based on syntactic or lexical restrictions which are sufficient to answer simple questions (such as factoid questions). Nevertheless,

semantically equivalent expressions cannot be detected by these lexical-syntactic patterns. In these cases, a deeper understanding of the question and the documents is necessary.



We consider that the detection of semantic relations could enable QA systems to find a greater set of candidate answers and to improve their recalls (proportion of correct answers retrieved). For this reason, we have participated in the task titled "Classification of Semantic Relations between Nominals" organized by the 4th International Workshop on Semantic Evaluation, co-located with ACL 2007.

In this task, a nominal is a noun or base noun phrase, excluding named entities. A base noun phrase is a noun and its premodifiers (nouns, adjetives, determiners). Complex noun phrases with attached prepositional phrases are not considered. The nominals can occur either on the phrase, clause or the sentence level. This fact constitutes the most important challenge in this task because the majority of the previous approaches were limited to the "compound nominals".

"Support Vector Machines, a machine learning algorithm, used to automatically classify semantic relations"

Organizers of SEMEVAL 2007 provided benchmark datasets which included seven semantic relatons: Cause-Effect (e.g., *virus-flu*), Instrument-User (e.g., *laser-print),* Product-Producer (e.g., *honey-bee*), Origin-Entity (e.g., *rye-whiskey*), Purpose-Tool (e.g., *soup-pot),* Part-Whole (e.g., *wheel-car),* Contain-Container (e.g., *apple-basket*). For each relation, the dataset contained 140 training sentences and 70 testing sentences. Each sentence is manually annotated with the following information: the nominal boundaries (e.g., *"<e1>Zinc</e1> is essential for <e2>growth</e2> and cell division*), the query used to retrieve the sentence from Web (*e.g., = "* is for growth*"*), the WordNet sense key for each nominal (*e.g., WordNet(e1) =* "zinc%1:27:00::", *WordNet(e2) = "growth%1:22:00::",*) and the value of the relation (e.g., *Cause-Effect(e1,e2) = "true"*). Approximately half of the sentences are positive samples and the other half are near-miss negative samples. Each relation is considered as a binary positive-negative classification problem.

Our system is based on the use of machine learning algorithms. Machine Learning, as a subfield of Artificial Intelligence develops programs which enable computers to acquire knowledge automatically from examples. In our case, a machine learning algorithm is used to classify automatically the semantic relations between nominals from training datasets. We decided to use Support Vector Machine (SVM), as it has achieved the best performances for many classification tasks such as document classification or semantic role labelling. In particular, we used Sequential Minimal Optimization (SMO) as a quick method to train SVM.

The set of features used for the classification of semantic relations includes information from different levels: word tokens, POS tags, verb lemmas, semantic information from WordNet, etc. The lexical and POS features are automatically extracted using the infrastructure for developing software components for Language Engineering GATE. Firstly, the sentences are processed using the English Tokenizer of GATE. A set of tokens is extracted in this stage. These tokens include: the nominals tokens, the two words before the first nominal, the two words after the second nominal and the word list in-between. These features are based on previous work addressing the detection of semantic relations between named-entities. In addition, sentences are processed by the Morphological analyser to extract words' lemmas. In order to

avoid data sparseness, sentences are labelled using the Part-Of-Speech (POS) tagger of GATE, and the POS tags of the above features are also considered. Moreover, POS tags are concatenated to create the feature *PATH*. We believe that the semantic relation could be strongly correlated to the information about the verb and the preposition. For this reason, the system detects if a verb or a preposition occurs in-between the two nominals and then features such as the token, voice and lemma in case of the verb, or token and type of the preposition are considered. The query provided for each sentence in the corpus is also considered.

Secondly, the system extracts semantic features from WordNet. A simple module obtains the synset number and the lexical file number for each nominal in the dataset. According to the WordNet architecture, synsets are organized into several semantic domains. In the case of noun synsets, there are 26 lexicographer classes or domains: act, animal, artefact, attribute, body, etc. We think that these classes can help to determine if the nominals satisfy the restrictions for each relation. For example, in the relation Theme-Tool, the theme should be an object, an event, a state of being, an agent, or a substance. Else, it is possible to assure that the relation is false.

"A semantic network of questions and documents could allow a logical inference mechanism"

The system is trained using 10-fold cross-validation. Cross validation is a method to analyze the ability of the algorithm to classify unseen examples and 10 is the fold number recommended when training dataset is small. The dataset is divided into 10 subsets. The algorithm uses all the 10 subsets except one to build a model to classify the examples of the remaining subset. This is performed through the entire data set, leaving one different subset each time. At the end, all the examples will have been used as validation set.

The system achieves its best performance using WordNet information. Our average score F is 64.3% and it is above the average of participating teams (F=63.6%) and the baseline. The best results are achieved for the relations: Instrument-Agency, Product-Producer, Part-Whole. However, for the relation Theme-Tool the system obtains lower scores.

These results represent the first stage of implementation. We believe that to improve the system performance, a wider use of semantic information is needed. For example, the immediate hypernym for each synset obtained from WordNet could help in improving the results. Besides, Named entities recognition could help in the detection of some relations like Origin-Entity or Product-Producer.

Furthermore, we would like to extend the system to deal with the semantic roles of verbs. Semantic Role Labeling aims at identifing predicate (verb) -argument structure within a sentence. This type of information could be very helpful in QA systems since in many cases the question 's main verb reflects the expected semantic relation in the answer. Semantic resources such as VerbNet, FrameNet and PropBank contain information about the predicate-arguments structures and could also be used.

The incorporation of information regarding the semantic relations between concepts as well as between the verb and its arguments into a QA system would allow a deeper analysis of the question and the documents and hence, a more precise set of candidate answers could be obtained. In addition, it would be possible to build a semantic network of questions and documents allowing a logical inference mechanism. This mechanism would enable complex questions.

We would like to thank Roxana Girju, Marti Heart, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney and Deniz Yuret for the Dataset provided for SEMEVAL 2007-Task#04, "Clasification of Semantic Relations between nominals".

Isabel Segura-Bedmar is a PhD student at University Carlos III of Madrid. Her research focuses on improving Information Retrieval and Question Answering Systems by the inclusion of Semantic Information and the use of machine learning techniques. She can be contacted via: isegura@inf.uc3m.es

Research Update:

Modeling and Annotation of the Dance Media Semantics

*By **Kannan Rajkumar***

Dance is one of the important cultural wealth of any nation. In ancient times, dancers passed the knowledge of the dance verbally to the following generations. Since human memory has limited capacity, many dance productions have been lost. Recently, technological advances in recording and storage media for digital information have made a rapid production of digital videos and audios. However, documenting the expressive semantics of the dance that exist in these digital collections and providing an efficient search facilities for the dance community such as dancers, dance students and choreographers remain a challenging task.

Project Aims

Dance data is essentially multimedia by nature consisting of visual (dance pieces), audio (music, tempo, intonation) and textual (lyrics of songs) information. My research represents the semantic based approach to the modeling, annotation, authoring and retrieval of dance video objects. The aim is to develop semantic models to incorporate the various dance video semantics of both Indian and generic dances, to allow dance experts to annotate the semantics, to perform the semi-automatic authoring of *MPEG-7* and *XML* instances from these annotations and to handle the dance users' semantic queries (www.informatik.uni-trier.de/~ley/db/indices/a-tree/r/Rajkumar:Kannan.html).

This research introduces the representational structures for the semantic (such as dancers, dance movements, context, culture, emotions etc), spatiotemporal (like relationships between the characters and their body parts) and relational features of the dance video; representations for narrative structures such as action and events; and representations and strategies for dance video queries.

These representations and strategies form the basis for the prototype systems, *IndVideo* (Indian dance videos), *DanVideo* (generic Dance Videos) and *DMAR* (Dance Media Annotation and Retrieval) which provide facilities for the semi-automatic annotation, authoring and dance semantics retrieval.

Applications

These dance video retrieval prototype systems have a wide range of applications, a number of which are explored in my research:

- *Dance Learning*: Dance students can learn the dance pieces of a dance concert themselves in the absence of the dancers who train the students by conducting the rehearsals.
- *Contemporary Dance Learning*: The annotation system archives several dances of multiple origins and cultures of several countries. Therefore, dance students can learn several dances from different origins by comparing the features of the dance pieces.
- *Self-paced Dance Learning*: Since dance student's learning capability varies, students can learn dance pieces at their own pace repetitively.
- *Culture Learning*: Properly annotated dance video is the cultural heritage to the present and future generations of the dance community.
- *Dance Training*: Dance training is a difficult and time consuming task for the dancers. Dancers can use these tools to train the dance pieces easily with less physical effort.
- *Choreography Design*: While designing new choreographic sequences, choreographers can consult the dance retrieval tool in order to avoid the duplication of the dance movements.

Metrics for dance annotation

The fundamental requirements of these dance video applications are:

- The generated annotations of the dance video features should be close to its original form.

- The retrieved results of the user's dance video queries should be effective.

For the first requirement listed above, I propose a metric, *fidelity* to evaluate the quality of annotations generated by these prototype systems. Fidelity is defined as the ratio between the number of dance pieces correctly resurrected or re-performed by the dancers by using the annotations and the total number of original dance pieces. The effectiveness factor is characterized in information systems literature, using the parameters *precision* and *recall*. Considering the second requirement, my research employs the *precision* and *recall* to characterize the extent to which the retrieved results are relevant to user's queries.

Conclusion

A central contribution of this research is the development of semantic models, annotation and authoring frameworks and query processing engines for the Indian and generic dance videos. The interaction with the dance teachers and students revealed their way of teaching and learning the dance pieces. As a result, it was found that the expertise of the dancers could not be fully codified in words, similar to other domains. This is because of the complexity involved in the training of a dance. That is, certain dance movements should be emulated, rather than explained. However, the physical presence of the dancers throughout the rehearsal session repetitively is a cumbersome task. But these are the important requirements of the dance students to learn the dance efficiently. In this research, some of the complexities of the annotation and retrieval tasks are reflected in our prototype systems.

Kannan Rajkumar holds BS, MS and M.Phil degrees in Computer Science from Bharathidasan University, Tiruchirappalli, India and is currently pursuing his Ph.D. degree in Computer Applications at National Institute of Technology, Tiruchirappalli, India. His research focuses on the Annotation, XML and Mpeg-7 Authoring and Retrieval of Dance Media Semantics. He can be contacted at cak0303 @nitt.edu and rajkumarkannan @yahoo.co.in.

Book Review:

"Multimedia Data Mining and Knowledge Discovery", Edited by Valery A. Petrushin, Latifur Khan

Reviewed by Andrew Neill

In *Multimedia Data Mining and Knowledge Discovery*, the editors have set themselves the difficult but pertinent task of assembling the leading, state-of-the-art research papers on what is one of the most public-facing developments in computing. One barely needs to read a newspaper, browse the internet or pick up a copy of *Wired* magazine without there being an article about YouTube, streaming IP-TV, podcasts, internet radio, online music sharing…the age of multimedia consumer applications over the internet has well and truly arrived. Add in applications for analysing security video, mining medical CAT scans and automatic iris recognition and you have a selection of current uses.

How, then, might you find what you are looking for amongst the multimedia riches on the internet? And how do we avoid having to sit and watch endless hours of video surveillance tapes to pick out wrongdoers? Potential answers to these are contained within this edited collection of academic papers.

Note, however, that there is no mention of these consumer applications within this book – no YouTube or Last.fm – this book is not about Web 2.0, not aimed at the "intelligent layperson". The book jacket states that it has been "written with graduate students in mind" – helpfully adding that "researchers with interests in multimedia data mining, summarization, indexing, and retrieval" will also find it a useful resource. Readers of Chapter 10 "*Visual Alphabets: Video Classification by End Users*" could draw a parallel with tagging YouTube or podcasts, but discussions of distributed colour histograms

and segmentation of HSI (hue, saturation and intensity) colour space will leave the non-technical behind. This book contains a selection of key research papers on multimedia analysis, data mining and knowledge management, describing experimental results, current challenges and future directions.

A brief glance at the list of contributors shows a long list of universities – and Accenture – across the USA and Europe, with a couple of contributors from Japan, Australia and China. Note that the papers were mostly taken from two conferences held in the US, so perhaps the cost of travelling to these events has unintentionally led to a geographical gap. A question – did this accident of geography lead to good research in South and East Asia being missed? Without being an expert in the field, it's difficult to tell, but I would have expected more from Japan, Korea, Singapore, China and India.

"This is useful in addressing such weighty questions as whether an algorithm can spot goals in a football match on TV"

The papers are grouped into five sections, with on average five papers per section. Papers are about twenty pages long, and are structured in the standard academic manner – introduction, overview of research, experiment, conclusions and future work. This simple structure allows for efficient browsing.

Part 1 is an introduction and overview of multimedia data mining. I found this section essential to refresh my memory of the topic from my Imperial College days (speaking of which, I'm disappointed not to see their multimedia team in the book!). Valery Petrushin has written an excellent and compact introduction on the first ten pages, with a particularly interesting summary of multimedia and data mining applications and a short explanation of each chapter. This is followed by an overview of data mining in slightly more technical detail by Nilesh Patel and Ishwar Sethi, which introduces a lot of the concepts and challenges. This first section provides the context for the rest of the book.

Part 2 is titled *Multimedia Data Exploration and Visualisation*, and deals with clustering of images and events, modelling of similarity, and analysis of motion capture. Part 3, which is incidentally the shortest, holds three papers about *Multimedia Data Indexing and Retrieval*, covering the semantic classification of video by examining the images within. Part 4 contains eight papers on Multimedia Data Modelling and Evaluation, which covers the tricky subject of computer-understanding – essentially, examining how well computers can recognise novel or interesting events within video, compared with a human's capabilities – determining what is interesting or new about a scene. This is useful in examining security video, or addressing such weighty questions as whether an algorithm can spot goals in a football match on TV (it is not clear whether the computer indicates this by leaping around the room with a shirt over its head).

"This is a wide-ranging and fascinating title... the quality is consistently high"

Part 5 provides some Case Studies and Applications. These are based on practical applications more than the earlier chapters. Paper cover the design of collaborative workspaces, extracting semantic meaning from video, evaluating the effectiveness of web banners, extracting user profiles by examining their use of a video search engine, and the laudable paper about data mining of medical images. This section interestingly also provides a hint of where this technology might lead, with papers about tracking people in an office environment (very Orwellian) and the automatic recognition of people using images of their iris which, if combined, closely approximate a paranoid science-fiction Hollywood film - think Tom Cruise desperately evading the authorities in Steven Spielberg's The Minority Report.

In conclusion, this is a wide-ranging and fascinating title, showcasing a lot of threads of research that will undoubtedly inform the world of the future. The choice of papers covers most of the major topics one would expect, and the quality is consistently high. However, it is difficult not to wonder whether

more recent conferences would be even more relevant (the papers were first presented in 2003 and 2004, although they have been updated and extended for publication). Also, there is a niggling doubt that this is a truly global selection, given the small representation from the Asian universities. Overall, however, this is a minor misgiving, and the book is highly recommended for researchers and computer science students with a keen interest in multimedia. Perhaps we will see the next YouTube from one of the contributors – or, perhaps more likely, one of its readers!

Andrew Neill MEng MSc MBCS is the Business Analysis, Research and Design Leader in the IT department for the international lawyers at Norton Rose. He specialises in web technologies, information retrieval and knowledge management, and has recently helped launch a new knowledge search engine in Norton Rose, based on the FAST Search platform. Future projects include maximising the value of the search engine, examining the impact of Web 2.0 on the legal business, and researching strategic IT applications. Previously, Andrew worked as a consultant at Deloitte & Touche in the advanced technology arm of their Business Consulting division. He is a graduate of both Imperial College and Strathclyde University, and lives with his wife in North London.

Obituary: Karen Spärck Jones

By Stephen Robertson

Karen Spärck Jones, who died from cancer on 4th April 2007, will be well-known to many members of the BCS IRSG. She was a commanding figure on the IR research scene, and simultaneously in the world of computational linguistics, in this country and around the world, for well over 40 years.

Karen's academic home for almost the entirety of her research life (after a period of research in Margaret Masterman's Cambridge Language Research Unit and also completing her PhD) was the Mathematical Laboratory, later Computer Laboratory, Cambridge's equivalent of a computer science department. But this was sometimes an uneasy home. IR has only relatively recently become a more-or-less mainstream part of computer science; NLP too was somewhat marginal to most views of CS. Her PhD thesis, about semantics and synonymy, must have seemed way out to those who thought of CS as programming languages, numerical methods and algorithm design. And one of her major early inspirations in IR was the now-canonical Cranfield experiment, conducted by a librarian without the use of a single computer. In fact you could say that initially she regarded the computer essentially as 'just' a tool for research. At around the same time, the publishers of large scientific abstracts journals were beginning to see computers as a suitable tool for their production work as well. This development led eventually to the first generation of computerised search systems, the predecessors of today's search engines.

Nowadays, of course, the computer is a universal, ubiquitous tool; if there is a need or a use for it in subject X, then you would expect an academic Department of X to have a computer or ten. But in the 1960s, the mountain would not come to you — you had to go to the mountain. That was exactly what Karen did, initially unofficially, around 1964 while completing her PhD. And although the

Computer Laboratory contrived largely to ignore her for many years, it remained her base of operations until she died. Her first official position there, but externally funded, was in 1968; she remained on 'soft' money for many years. Eventually the University gave her the recognition she had long deserved: she was appointed to a Readership in 1994 and to a personal chair in 1999.

In the meantime, she had become one of the most active researchers in IR; put it somewhat to one side for a while to concentrate on NLP; and then returned to the field, more-or-less with the arrival of TREC, which was in substantial measure inspired by her earlier work.

If I mention here some of her research activities and achievements, it is in the consciousness that it is really hard to do justice to them, to give each of them fair weight. But at this point some such list is needed, so here goes.

Before she even started her PhD, a paper (authors Masterman, Needham and Spärck Jones) was given at the famous 1958 conference on scientific information. The paper's title, "the analogy between mechanical translation and library retrieval", clearly indicates the areas that interested her. Her subsequent PhD work on synonymy led to ideas on the use of term clustering methods (specifically those based on the theory of clumps, developed in part by Roger Needham). She spent some years, supported for a while by a Newnham Research Fellowship and then by a Royal Society Research Fellowship, in pursuing this line, and acquired the Cranfield collection in the process. Although the clustering work was not very successful, it led her on in two notable directions. One of these was term weighting; she developed the term weighting formula that subsequently became known as idf, and then she and I worked on the relevance weighting model -- both successful and widely influential.

The other direction was a concern with experimental work in IR and the development of useful test corpora. Having never built her own test collection, but instead relying on others built (like Cranfield) for specific purposes, she was more aware than most at the time of the necessity of constructing well-designed corpora. The work in the late 1970s on the so-called 'ideal' test collection was seminal. But it seemed ahead of its time — the UK research funding available at the time was not really able to support such a substantial effort, and the plan went on the shelf.

She then edited a collection of papers, published in 1981 as the book Information Retrieval Experiment, on the methods and principles of experimental research in the field, as well as some of the results. This and the ideal test collection reports represent the first serious attempt, outside projects built around specific experiments, to establish a foundation of good experimental practice and methods for the field.

Actually, to say simply that she 'edited' the book does not quite do justice to her approach. She was never one to sit back and let the authors do their own thing; she was always actively engaged in the process, making sure that authors fulfilled their commitments and worked together where necessary, and that everything in the book satisfied her own exacting standards.

However, the apparent failure of the ideal test collection initiative left her somewhat dispirited about the IR field, and for a while in the 80s she concentrated much more on her NLP work. This led to service as President of the Association for Computational Linguistics in 1994, an active and radical presidency.

At the beginning of the 1990s, TREC was announced. TREC, the Text REtrieval Conference, was the ideal test collection writ large, and underpinned by US funding. Suddenly, experimental information retrieval research took on a new lease of life. Although Karen did not do much actual experimental work in TREC (she entered just a handful of tasks herself in a few early TRECs), she served on the Programme Committee throughout, and also joined as an advisor to my team at City University. (The BM25 ranking function which we developed in the early 90s and which has now acquired a kind of canonical status in the field is a direct descendant of the relevance weighting model of the 70s.) Most importantly, she contributed a series of great

'Reflections on TREC' pieces which succeeded admirably in capturing the lessons we had variously learnt from TREC, otherwise deeply buried in the rather impenetrable reports of participants, or else emerging and disappearing in the discussions at the conference itself.

Throughout this time, she was fully engaged in whatever details of the research field took her fancy. She masterminded a further series of experiments (not part of TREC but making extensive use of TREC data -- as she had envisaged in the 70s) on the probabilistic relevance models. She plunged into discussions on many newly emerging ideas such as the language model approach to IR. She also pursued her NLP interests, playing a major role in the development of the TREC-like Document Understanding Conference (DUC), which is aimed at summarising tasks.

Karen married Roger Needham in 1958, a long and happy marriage ended untimely by Roger's death from cancer in 2003. She was also active in many other spheres, most particularly in encouraging and promoting the cause of women in computer and information science. She had a huge enthusiasm for life, which will be sorely missed.

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.