Informer

Number 24, Autumn 2007
ISSN 0950-4974

BCS INFORMATION RETRIEVAL

In This Issue

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.

So - communities, then. Who needs them? All of us, it would seem – whether it's on Facebook, Twitter or some other social networking phenomenon, you can't have escaped the hype.

But that's the point – this is more than just hype. Most of these sites succeed because they fulfill a genuine need – providing a forum for people to interact, exchange information and ideas, and most importantly, build relationships.

Which is why, on occasion (well, quite often actually) I go on a bit of a rant about the IRSG. Am I the only one that finds it acutely ironic that a group constituted to advance the management of information within society at large seems so unable to manage the information and ideas of its own members?

I am, of course, referring to online communities - and our own lack of such a facility. Yes, I know we have various mailing lists and so on – but a group where 90% of the traffic is conference announcements isn't my idea of a community.

So that's why I was pleasantly surprised to learn about SearchCoP – a Yahoo group set up by Seth Earley and Avi Rappoport, which describes itself as a "*forum to share ideas, techniques and experiences in designing, configuring, testing, analyzing and maintaining enterprise search engines and search applications*". As with any community, it will take time and effort to build, but I am hopeful that in the longer term this group will provide a valuable forum for our practitioner members, who I think are perhaps the group least well served by our current portfolio of activities, events and publications.

Meanwhile, onto this issue of Informer. I'm pleased to present the second half of our series of Research Updates, submitted by attendees at the Summer School on Multimedia Semantics. We have four excellent articles, again showing the breadth of ideas and creativity in our community.

We should also show our appreciation to *Christina Lioma*, who contributes an excellent review of "*Natural Language Processing and Text Mining*", by Kao and Poteet. By the way, we still 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.

All the best,
Tony

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

Flag and Bell Tech Networking Event

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.

The FAB Christmas Special in December will be sponsored by Skype. We will meet at Skype's office from 6pm onwards before moving to the first pub.

Date: Tuesday, 4th December 2007
Address:
Skype Office
2 Stephen Street
London W1T 1AN

Map:
http://streetmap.co.uk/newmap.srf?x=529675&y=181575&z=0&sv=W1T+1AN&st=2&pc=W1T+1AN&mapp=newmap.srf&searchp=newsearch.srf

Closest tubes: Tottenham Court Rd and Goodge Street

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

BCS-IRSG Announcements from the Chair

by Leif Azzopardi

Greetings!

During the last week of August, the first BCS-IRSG symposium on Future Directions in Information Access (FDIA 2007) was held during the European Summer School in Information Retrieval, at the University of Glasgow. The symposium ran over two days and was not only great success with over 60 students attending along with over 25 leading IR academics, but also great fun too, with lively sessions and a good night out with lots of food and wine!

The aim of the symposium was to provide a forum where future directions of information access could be presented and discussed in an open and friendly environment. To the name of the symposium included, "Future Directions" because we wanted to encourage submissions that focused on the possible paths and directions of work to be done to stimulate discussed and interaction. And the focus was on "Information Access" instead of "Information Retrieval", because it captures the broader notions of information retrieval, storage and management to include interaction, visualization, usage, and beyond.

As part of the programme, 18 students presented their work in fast paced five minute presentations to advertise their research, before a breakout poster session allowed the audience and presenters to discuss their research in detail. By all accounts presenters found this a great way to get really detailed feedback from more experienced researchers and their peers. Afterwards, we all went to the Primary (a local pub) for several well earned drinks, some food and lots of chat, games, and music.

On the second day, I presented the set of draft ECIR guidelines for writing IR papers, followed by additional guidelines for evaluation and multimedia papers from Steve Roberston and Arjen de Vries, respectively. These guidelines provide really useful tips for writing papers and point out some of the implicit questions that a reader expects to be answered when reading a paper. They are a good starting point and reference point especially when writing a paper that isn't just purely experimental. Afterwards, Theo Huibers, presented a business case for IR in the style more like a stand up comedian than a professor, which was thoroughly enjoyable and entertaining. The day concluded with a panel chaired by Prof Keith van Rijsbergen, on "what makes a good PhD in IR", which was of great interest and value to all the students.

On behalf of the BCS-IRSG, I would like to thank all the Speakers, Panelists, Programme Committee and other volunteers who helped make the symposium a success.

SSMS 2007 and ESSIR 2007: In July and August, the BCS-IRSG sponsored two summer schools held in Glasgow, the first was the Summer School on Multimedia Semantics Analysis, Annotation, Retrieval and Applications and the second was the European Summer School on Information Retrieval. We provided travel grants to 24 students to attend these events, and during this time we attracted over 150 new members to the BCS-IRSG who also attended these events. So welcome to all our new members and I hope to see you at a future BCS-IRSG event soon!

ECIR 2008: Speaking of future IR events, the BCS-IRSG European Conference in IR is also being held in Glasgow organized by Iadh Ounis as General Chair from the University of Glasgow and Ian Ruthven as PC Chair from the University of Strathclyde. The paper and poster submission deadlines have now passed, and now it is an anxious wait for well over 200 authors till the notification date in December. ECIR 2008 will be in its 30th year, and will include workshops, tutorials and an industry day along with the main conference. It will be the biggest ECIR yet!

Search Solutions: Next year we plan to run the second Search Solutions event (SSE 2008) in September, 2008. This event follows on from Search Solutions (SSE 2007) held in May, 2007 and is devoted to the interests and needs of information retrieval practitioners and are aimed at exploring the challenges involved in designing and developing practical solutions for information search and retrieval. A specific goal of Search Solutions is to build bridges between IR specialists in industry and academia. So this forum presents an opportunity for commercial organisations to share their work with a wider audience, and for researchers to learn more about the issues faced by search professionals in developing innovative search and retrieval products and services. Pictures and Talks from SSE 2007 are now available from the BCS-IRSG Website.

Other events in 2008: Also in September, will we run the second symposium in Future Directions in Information Access. Then in October, the BCS-IRSG is organizing in conjunction with Queen Mary University the 2nd symposium on Information Interaction in Context. For those who live in or near London, you'll be happy to know that all these events will be held at the BCS-HQ in Covent Garden.

Committee Members Wanted: Just to let you know now, at the next AGM, we will be looking for new committee members to come on board and help organize more IR events in the UK and abroad. If you are interested in joining the committee, please get in contact with myself or Andy MacFarlane (the IRSG Secretary) to discuss how you can get involved.

ECIR 2009: Finally, in response to the call for bids to host ECIR 2009 we received three high quality bids, which made the decision particularly difficult. After much deliberation and discussion the BCS-IRSG committee decided to award the hosting of the 31st BCS-IRSG European Conference in Information Retrieval to the Toulouse bid organized by IRIT and the University Paul Sabatier Toulouse with Prof. Mohand Boughanem as General Chair.

Leif Azzopardi is an RCUK Research Fellow in the Information Retrieval Group at the University of Glasgow. His research interests include: the behaviour of information retrieval models, the study and evaluation of information access systems, and the theoretical development of formal information retrieval models. He can be contacted by email via: leif@dcs.gla.ac.uk.

Research Update:
New Light on Information Agents

By Milan Stankovic and Uros Krcadinac

Agent-based systems are one of the most important and exciting areas of research and development that emerged in information technology (IT) in the past two decades. In a nutshell, an agent is a computer program that is capable of performing a flexible, autonomous action in typically dynamic and unpredictable domains. In particular, an information agent is an agent that may access one or multiple, distributed, and heterogeneous information sources available, and pro-actively acquires, mediates, and maintains relevant information on behalf of its user(s) or other agents, preferably just-in-time.

EMERGING TRENDS

Information agents emerged as a response to large and ever-increasing quantities of available data. Blogs, Wikis, and other social software have added a new value to the Web by making it a source of information on various topics. New content is being added continuously by users who actively publish new texts and multimedia.

Many Web users express need for getting relevant information from various Web sites, but do not have time to inspect them in regular basis. It is fair to say that RSS and many available RSS readers enabled users to easily track changes on constantly changing web resources, and new tools, like Yahoo Pipes, enabled them to filter and combine sources in order to retrieve information of their interest. However, there are many sites that do not provide a RSS feed, and among them there are sites that contain valuable information.

INFORMATION GATHERING MULTI-AGENT SYSTEM

In order to solve this common problem of user-oriented information gathering, our research group at the Department of Information Systems, FON - School of Business Administration, University of Belgrade, Serbia, has developed a multi-agent system, called Talaria System (http://iis.fon.bg.ac.yu/talaria/). Talaria System (The Autonomous Lookup And Report Internet Agent System), named after the Greek Messenger God Hermes's winged sandals, is a multi-agent system that integrates information gathering and filtering in the context of supporting a user to manage her/his Web interests.

The system provides each user with a personal agent, which periodically monitors the Web sites that the user expressed interest in. The agent informs its user about relevant changes, filtered by assumed user preferences and default relevance factors. One of the most important characteristics of this agent-based software are its usability advantages. Human-agent interaction is completely realized via email, so that a user can converse with her/his agent in plain natural language, whereas the agent heuristically interprets concrete instructions from the mail text (e.g., "change site list" or "kill yourself"). In order to improve characteristics of Web sites which do not provide RSS feeds, one of the features of the Talaria System is the creation of RSS feeds based on the new and presumably relevant content of the Web sites that are being inspected.

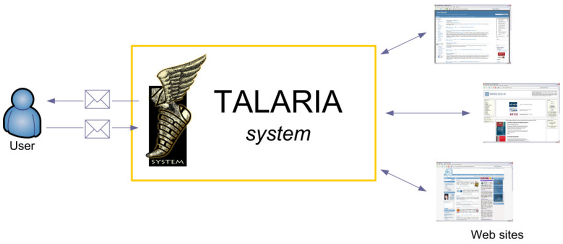


Figure 1 - TALARIA System

Talaria System is being mainly tested for academic proposes, inspecting Web sites of various faculty subjects and informing users about possible changes of schedules and exam dates, and it proved to be of vital importance to students. Human-like NLP-based interaction, autonomy-related aspects of this system, and acting on behalf of the user emphasize the usability advantages of this agent-based software.

Conclusion

Concerning the fast development of the World Wide Web, it is clear that there is an increasing need for information gathering services, in order to help Web users find the right information just-in-time. We have developed our system in order to meet some of those requirements, but it is just a small step, and there is a lot of research to be done. Future research in the field of information agents must take into an account all major IT trends, such as the Semantic Web, Web services, grid computing, autonomous systems, and many more. It is our opinion that all of these visions share the common need for agent-based technologies, and that information agents (or something very similar to them) will continue to be one of the most vivid areas of the research in the years to come.

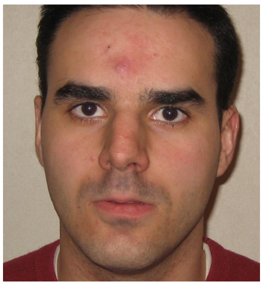
Uros Krcadinac is an undergraduate teaching assistant at the Department of Information Systems and Technologies, FON - School of Business Administration, University of Belgrade, Serbia. He is a member of the GOOD OLD AI research group (http://goodoldai.org.yu/). His research interests include Intelligent Agents, Multimedia and Semantic Retrieval. He can be contacted at uros@krcadinac.com.

Milan Stankovic is an undergraduate teaching assistant at the Department of Information Systems and Technologies, FON - School of Business Administration, University of Belgrade, Serbia. He is a member of the GOOD OLD AI research group (http://goodoldai.org.yu/). His research interests include Intelligent Agents, Information Retrieval, and Social Software. He can be contacted via milan.stankovic@gmail.com.

Research Update:

Scalable semantic personalised search of spoken and written contents on the Semantic Web

By Javier Tejedor



The continuous growth of contents in text and other media on the Web and intranets demands a parallel progress of information access technologies. In this project we propose:

(a) the development of a novel semantic search model, with ontology-based content ranking algorithms, improving and complementing the achievements in this direction in the Semantic Web field, and achieving measurable improvements in search effectiveness with respect to current keyword-based techniques;

(b) the combination of personalisation techniques with the semantic-based representation models, to achieve improvements in the relative precision and relevance of search results with respect to the particular interests of individual users;

(c) the integration of text and voice contents in a single access platform for large-scale repositories; and

(d) an experience in the realisation of Semantic Web proposals, to contribute to its advancement, including the definition of ontologies, the semi-automatic construction of knowledge bases, and a semi-automatic content annotation system, incorporating speech recognition and human language technologies.

A demonstrative application will be built in the news domain, for the evaluation and refinement of the developed techniques, in a real-world application setting. In this article, I will give a global vision of the project modules and an overview of the most important speech recognition techniques applied nowadays and within this project. Finally a conclusion will be also presented.

GLOBAL VISION OF THE PROJECT MODULES

The scenery of the project will be as follows: A user makes a query. This query is passed over a semantic query processing in which it is executed against a knowledge base (KB). In the KB, the concepts related to the application domain are formally described by means of ontologies. The concepts which satisfy the query are retrieved by this process and serve as an index to search for the keywords in an XML file annotated previously from the speech-based techniques. The concept-keyword mapping makes possible to transform the concepts in the ontologies into keywords in the speech recognition vocabulary. Finally, a ranking strategy proposed to order the documents where the keywords appear will allow us to show the user the final output, in order to he can listen to the speech content and read the text documents according to his/her query. A friendly user interface is also being investigated within this project. The whole architecture can be seen in the next Figure 1:

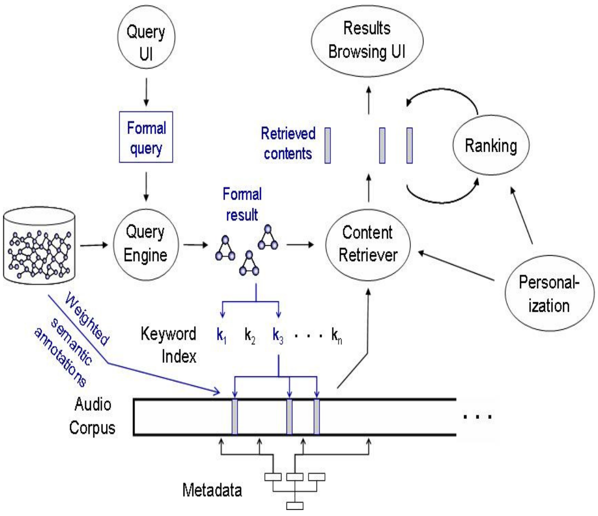


Figure 1. Global system architecture

More information about this project can be found in (http://nets.ii.uam.es/s5t).

SPEECH RELATED TECHNIQUES

As it is well known, dealing with the audio content implies the development of speech technology-based techniques in order to find relevant keywords contained within it. Normally, Large Vocabulary Continuous Speech Recognition (LVCSR) systems outperform any other kind of techniques, but this approach has three main drawbacks, a) it is necessary a very high amount of data in order to train the system, b) it has a high computationally cost and c) it is vocabulary dependant. In addition to this, a high knowledge about the language is needed. In this project, not only LVCSR-based techniques are going to be applied but also keyword spotting-based techniques are being investigated in order to search for a reduced set of keywords which commonly do not appear in the vocabulary of LVCSR systems such as proper names, acronyms, etc.

These techniques take advantage of the little amount of data necessary to train the system and of a faster recognition process, as well as may be applied over an opened vocabulary in contrast with the LVCSR system. Most of the keyword spotters developed for years used filler or garbage models in order to deal with the non-keyword segments (the Out-Of-Vocabulary (OOV) words) which appear in the speech. These filler models vary from phones to whole words. In this project, our filler models are composed by phones, phonemes and classes phonemes and no word models garbage are required. Hidden Markov Models (HMM) are used as acoustic modelling and HTK toolkit is used to perform the recognition processes. New confidence measures based on the scores of the recognizer output and on a lexical access module to build the keywords from the sequence of phones in the output of a phone recognition process are being used in order to achieve the best possible system performance and a good compromise between correct keywords and false acceptances. The architecture which achieves the best results nowadays in the keyword spotting system is shown in next Figure 2:

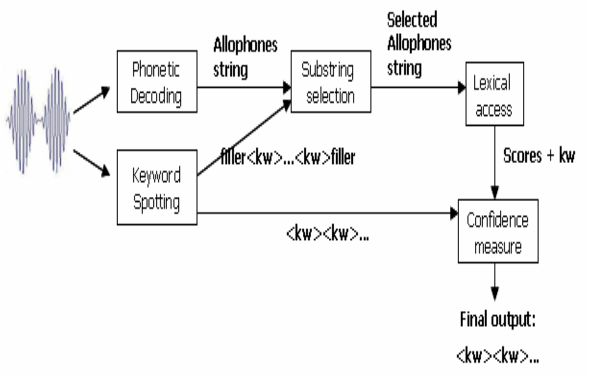


Figure 2. Keyword Spotting system architecture

The approach in this architecture performs a Keyword Spotting process to retrieve the sequence of keywords and filler models. At the same time, the Phonetic Decoding retrieves the allophones string related to the speech files. The Substring selection extracts the allophones string in the time intervals where the keyword spotting proposes a keyword. This string is passed through a Lexical access module, which, from a previous confusion matrix trained from the substitution, insertion and deletion allophones errors in this string and from a dynamic programming algorithm proposes the most probable keyword related to this string. The final Confidence measure eliminates from the keywords proposed by the keyword spotting module those which do not coincide with the one proposed by the lexical access and those whose score retrieved by the lexical access is higher than a threshold. Both the threshold and the confusion matrix were trained from a development set, different from the one used in the system evaluation.

By other hand, normally the sound files recorded from the radio and television programs, with which we have to deal within this project, do not only contain clean speech but also noises, music, etc. In this way, there will be lots of samples in which phenomena such as overlapping, noises, breathes and other effects occur. It causes errors in the recognition process. Therefore, the project will also investigate speech segmentation techniques and robust speech recognition techniques, e.g. those which try to separate the clean speech of the noise, in order to improve the search of the keywords within this news domain.

CONCLUSIONS

In this project we aim to develop a system which allows the user to make a query and retrieve the text and sound files which satisfy his/her query. In this way, semantic web area by means of ontologies, speech recognition by means of keyword spotting and LVCSR systems and speech segmentation are going to be investigated within it. It will also be investigated the user-friendly interface over which the final results will be presented to the user.

As preliminary results, we have presented an architecture for the keyword spotting system in http://jth2006.unizar.es/finals/4jth_106.pdf.

Update research outperforms the rate achieved with that architecture and presents a 93% of keywords (concepts) detected correctly and 11% of false acceptances with the architecture presented in Figure 2, over 80 keywords appearing 1672 times extracted from the clean speech Spanish Albayzin database test geographic corpus.

Javier Tejedor Noguerales is a 28-old PhD student in the University Autonoma in Madrid (UAM). I am also an assistant professor in that university in the "Theory of signal and communications" department. I belong to Human Computer Technology Laboratory (HCTLab) and my main research is focused on information retrieval of audio content (Audio mining) and my PhD-thesis is based on it. The contact e-mail is: javier.tejedor@uam.es

Research Update:

An approach to the extraction of chronological information: Temporal Expression Recognition and Normalization in Spanish

By María Teresa Vicente-Díez

Temporal information extraction from multimedia data sources supposes important linguistic challenges at the time of identifying and representing temporal expressions. For instance, it is frequently needed dealing with big amounts of information obtained from news and electronic documents. Since it is usual that this kind of documents is scarcely temporally annotated, it is difficult to be able to determine the time when the events of their contents occur. Dating of contents is crucial in the Information Retrieval (IR) field. The extraction of chronological information allows improving the results obtained by typical IR applications, such as Question Answering (QA) or Automatic Text Summarization. In QA it is fundamental to solve references that can help finding the answer to temporary questions ("*What year did Mozart die?*") or to questions with time restrictions ("*Who was the president of the USA in 2005?*").

One of the aims of my PhD research is the development of technologies for the extraction of temporal information from different data sources, in order to apply them to event ordering of the contents, as well as to the integration of a time reasoning system in a QA environment. In such a platform, it is expected that the introduction of inference rules allows the improvement of the question analysis as well as the quality of the extracted answers.

Spanish Resources

The scientific community counts with several resources in the field of temporal expressions management but, in their majority, only the English language is considered. Among them, there are several guides and methods of annotation, as well as temporally annotated corpora. Since Spanish is one of the most spoken languages in the world, it seems appropriated to invest in research and development of resources and systems for this language.

TERN Task

Temporal Expressions Recognition and Normalization (TERN) systems process input data from documents in a specific language (in this case Spanish) identifying dates, durations, anchors and intervals in them. The processing of each document includes several stages as is depicted in Figure 1, from the pre-processing of the source data to the return of the results in a standard format.

The TERN system in which I am currently working on identifies dates and intervals in a ***Recognition phase*** carried out by a *Recognizer* module. Both absolute (*April 3, 2005*) and deictic (*yesterday*) expressions are detected. Afterwards, this system manages and returns them in a standard format that avoids semantic ambiguities in their retrieval (***Normalization phase***).

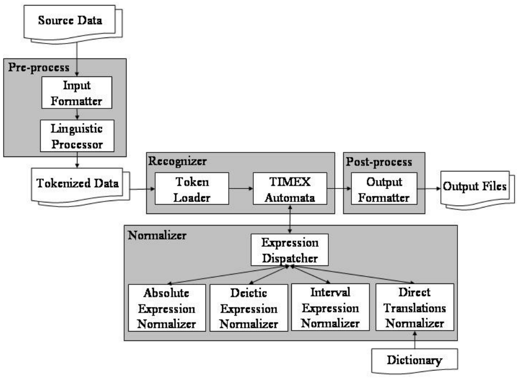


Figure 1 A TERN System Architecture Proposal

Since there are different types of temporal expressions, each one (absolute, deictic, intervals or those that directly suppose a point in time (*i.e.: Christmas, Halloween*)), has to be treated independently. The *Normalizer* module takes into account this fact, so it counts on a *Dispatcher* sub-module that identifies the different cases and deals with them in a proper way. An example of this classification, as well as the obtained normalized output is presented in Table 1.

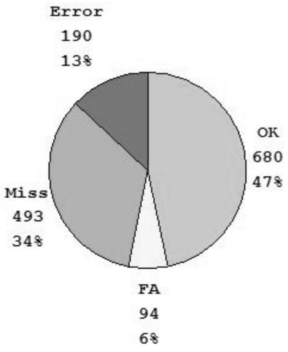
Table 1: An Example of the Temporal Expressions Classification Suggested

EXP. TYPE	**INPUT FORMAT**	**INPUT EXEMPLE**	**NORMALIZED OUTPUT**
ABSOLUTE	DD-MM-YYYY	31-12-2005	2005-12-31
	[DET]+DD+”de” +MONTH+”de” +YYYY	[el] 31 de diciembre de 2005 (*[the] 31th of December of 2005*)	2005-12-31
	MONTH + “of” + YYYY	diciembre de 2005 (*December of 2005*)	2005-12
DEICTIC	PRESENT_REFERENCE	hoy (*today*)	2007-06-01
	MODAL VERB “HACER” + QUANTITY + TIME_UNIT	hace un mes (*one month ago*)	2007-05-01

This proposal was presented to the *Automatic Content Extraction (ACE) 2007 Evaluation Workshop*, participating in the TERN task for Spanish language. The task using Spanish data supposed a pilot experience this year. This work constitutes a preliminary approximation where simple manual rule-based techniques are implemented both in recognition and normalization phases. The system focuses on the treatment of temporary expressions frequently appearing in the Spanish language, which should be labelled according to *TIDES 2005 Standard for the Annotation of Temporal Expressions* [http://timex2.mitre.org/annotation_guidelines/2005_timex2_standard_v1.1.pdf]

Testing the system

Some runs with several corpora, composed of news in Spanish, were made in order to evaluate the system. These results gave us a global estimation of the amount of expressions detected as well as the quality of the marked detections. Results obtained are considered quite successful since the number of *spurious* expressions (also know as ***false alarms***) detected by the system represents a low percentage of the total detections made. In the same way, the total number of ***missing*** and *misrecognized* (***errors***) objects is quite acceptable in the majority of the documents analyzed.



Conclusions

Completely integrating the TERN system into a QA system is the next step that must be done in order to check the improvement that it supposes at the time of the IR.

In the meanwhile, and in order to enlarge the number of temporal expressions properly recognized by the system the first enhancement that should be carried out is the tagging of the recognized expressions (durations, sets of times, etc) in order to capture as much semantic information as possible. Likewise, the grammar coverage for the recognition of temporal expressions that is considered by the system might be increased, adding other types of expressions that are not currently taken into account. Moreover, future work includes the implementation of dictionaries with a broader coverage of directly translatable temporal expressions, such as holidays, festivities, etc.

In the same way, it is high-priority considered the research concerning context extraction mechanisms that facilitate the recognition of deictic temporal expressions.

Finally, we are studying the introduction of machine learning techniques in temporal expressions recognizing and classifying stages, so that they complement the current rule-based approach.

María Teresa Vicente-Díez is a PhD student at the Advanced Databases Group of the University Carlos III of Madrid. Her research in the field of Natural Language Processing is focused on the recognition of temporal expressions, and the resolution of their temporal reference, applied to event ordering and time reasoning, with the aim of improving Question Answering Systems. She can be contacted by e-mail via: *teresa.vicente@uc3m.es*

Research Update:

Enriched Multimedia Ontologies by using fuzzy concepts

By Giuseppe Serra

The rapid increase of the available amount of multimedia information has revealed an urgent need for developing efficient automated multimedia information retrieval algorithms. In fact user increasingly desire to search on complex queries such as "Show me the video clip where both George Bush and Tony Blair enter in a car" or "Show me the video clip where Maradona bounces a ball".

Moreover such annotation would not only help human user to query and manage their digital libraries, but also enable automated applications performing complicated tasks like video surveillance to create, store, exchange and reason with the data.

For this reason research community has focused on the challenging problem of the semantic gap, the gap between the information that can be extracted automatically for visual data and the interpretation that the same data.

Significant progress has been made on automatic segmentation or structuring of multimedia content and the extraction of low-level feature. However, comparatively little progress has been made on interpretation and generation of semantic descriptions of visual information.

In the last years researches have used ontologies in order to achieve semantic analysis.

My research group and I will focus on our search in two main directions: creation of enriched ontology with video concepts and video event representations; study of fuzzy logic application to video concepts.

Multimedia Ontologies

Ontologies are formal, explicit specifications of a domain knowledge: they consist of concepts, concepts properties, and relationships between concepts and are typically represented using linguistic term. Ontologies can effectively be used to perform semantic annotation of multimedia content. In the case of video annotation the terms of the ontologies are associated to the individual elements of the video either manually of automatically, exploiting the result of the advancements in pattern recognition and image / video analysis.

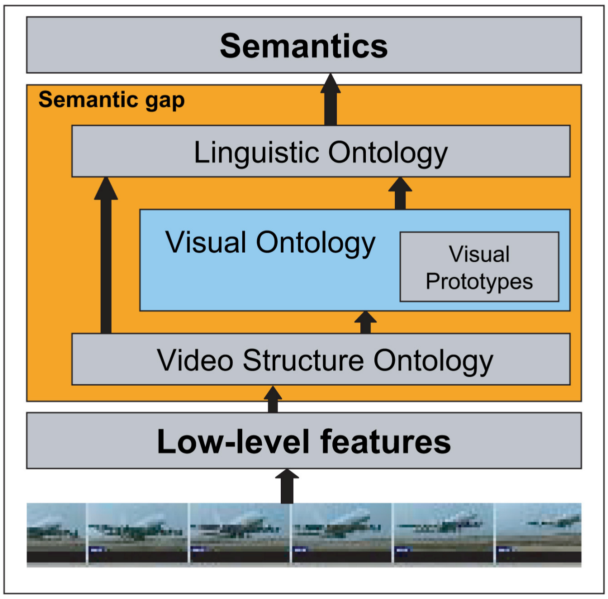


Figure 2 Several levels of ontologies are used to bridge the semantic gap between data and semantics.

Although linguistic terms, commonly used to define concepts in ontologies, are appropriate to distinguish event and object categories, they are inadequate when they must describe specific patterns of events or video entities. Consider for example the many different way in which an attack action can occur in soccer; these actions could be expressed in linguistic term only with a very complex sentence.

This motives the possibility that events that share the same patterns can be represented by visual prototypes instead of linguistic terms.

Fuzzy Logic

In a video the detection of the concepts can be have difference confidence degrees. This could be due to the similarities shared among different objects and the different appearances an object may have, so it is hardly possible to reach unique reliable annotation. For this reason the fuzzy logic is introduced in order to change the confidence degree of detected concepts analyzing their relationships. As an example, an image of a cow is more likely to be present in a landscape environment, like a

green field, whereas a desk is usually found indoors.
Moreover the fuzzy logic can be used to inference new fuzzy concepts using two or more detected concepts determining the confidence degree of the new fuzzy concept based on the confidence degree of the detected concepts.

Conclusions

In this paper we have briefly described two different topic, Multimedia Ontologies and Fuzzy Logic. In the future my group and I want to use the Fuzzy Logic in the Multimedia Ontologies; namely insert the degree of confidence for each instance of every concepts, (including the visual concepts). In this way we'll enable to apply the fuzzy reason theory to determine new inference fuzzy complex concepts.

Giuseppe Serra is a Ph.D. student MICC laboratory at the University of Florence, Italy. His research interests focus on video understanding based on statistical pattern recognition and ontologies and multiple view geometry. He was a visiting student at Carnegie Mellon University, USA in 2005 where he worked on Self-calibration Field. He can be contacted via serra@dsi.unifi.it

Book Review:

"Natural Language Processing and Text Mining", Edited by Anne Kao and Stephen R. Poteet

Reviewed by Christina Lioma

This book is a collection of articles on Natural Language Processing and Text Mining, both of which are currently enjoying a surge of interest fueled by the popularity of the Internet. Text Mining is generally defined as the discovery and extraction of high-quality information from text. The architecture of a Text Mining system typically involves structuring the input text (e.g. parsing), deriving phenomenological units of information (e.g., patterns) from it, and finally evaluating and interpreting the output. Natural Language Processing (NLP), the use of linguistic features to automatically extract a fuller meaning representation from text, is closely related to each of these processes.

This book, edited by Kao and Poteet, has resulted from a panel discussion at the 2004 ACM SIGKDD (Special Interest Group on Knowledge Discovery and Data Mining) Conference, which asked how Text Mining and NLP can best interact with each other. The editors have selected 11 papers, which cover a range of typical Text Mining tasks, such as text classification and clustering, concept/entity extraction, sentiment analysis, entity relation modeling, and retrieval.

Thematically, the book can be split into three main parts: NLP applications (Chapters 2-5), non-NLP applications (Chapters 6-9), and overviews (Chapters 10-11). Chapter 12 is a technical tutorial on textual processing, which addresses the needs of a very specific readership, and could perhaps have been included as an appendix.

The first part comprises four chapters, each treating a different aspect of NLP for Text Mining. In particular, the areas covered are opinion mining (Popescu and Etzioni), entity relation extraction (Bunescu and Mooney), domain-specific task classification (Mustafaraj et al.), and NLP-based Web search (Marchisio et al.).

The contributions are very interesting; each of the four areas covered are the object of much real-world research at the moment. Additionally, each chapter is presented and discussed in a coherent way that makes it easy to follow. Overall, the focus of these chapters is on the implementation aspect of the work presented.

The second part of the book is dedicated to non-NLP applications for Text Mining. It comprises four chapters on descriptive and inferential statistical methods. The work reported in these chapters makes use some light NLP, however the main focus is on applied statistics. The chapters cover Latent Semantic Analysis (LSA) for computer-assisted learning (Boonthum et al.), LSA for textual signature identification (McCarthy et al.), Support Vector Machines (SVM) for domain-specific document separation (Schmidtler and Amtrup), and Genetic Algorithms (GA) for enhanced text mining (Atkinson). Again, these chapters are interesting, well-written and easy to read. In this part of the book some novel work is presented, such as Atkinson's very interesting application of Genetic Algorithms to Text Mining.

The third part of the book includes two overview papers, namely an exploration of term weighting for classification (Liu et al.), and a survey of automatic ontology evaluation techniques (Brank et al.). Both studies have good coverage, and also include an evaluation component. Investigations of this kind are generally welcome in the field, because they provide a sense of research context. In this book, these two chapters may serve as reference points for the preceding work.

Overall, this book constitutes a nice contribution to the current literature in the field. The parallel investigation into NLP and Text Mining techniques and applications, which is the focus of the book, is of growing interest to several communities, Information Retrieval included. After all, Information Retrieval and Text Mining are both concerned with extracting information. In this book, emphasis has been given to practical applications of real-world problems, and perhaps less to the theoretical grounding and implications of the applications presented. According to the editors, this is an intentional decision, so that the work presented is not the type to be often found in conference proceedings, but rather algorithmic breakthroughs. Consequently, the presentation of algorithms and techniques is quite thorough in this book.

In summary, this book offers a balanced selection of contributions on a range of different Text Mining tasks. The chapters are coherent, and present mostly hybrid approaches that combine NLP and/or applied statistics. The book makes a pleasant read and is of interest, not only to the Text Mining and NLP communities, but also to IR practitioners.

Christina Lioma is a Ph.D candidate with the Information Retrieval Group, in the Department of Computing Science at the University of Glasgow. She holds a M.A. (Hons) in Languages and Linguistics from the University of Glasgow, and a M.Sc. (with Distinction) in Natural Language Processing from the University of Manchester. Her research is mainly on Computational and Quantitative Linguistics for textual Information Retrieval.

Forthcoming Events

Edited By Andy MacFarlane

Digital Media: Content and Communication Track at Hawaii International Conference of System Sciences (HICSS)
Ever worried about the 'digital tsunami'? This is the place to find out about it. Waikoloa, Big Island, Hawaii, 7-10 January 2008.
http://knowledgediscoveryconference.org/

Second International Workshop on Cross Lingual Information Access, Addressing the Information Need of Multilingual Societies (IJCNLP-2008)
A cross language information retrieval workshop.
Hyderabad, India, January 11, 2008.
http://search.iiit.ac.in/CLIA2008/

Fourth Asia Information Retrieval Symposium (AIRS 2008)
An Asian conference in the general area of IR, including NLP, evaluation, theories etc.
Harbin, China, January-16-18, 2008.
http://ir.hit.edu.cn/airs2008

International Workshop on "Metadata Mining for Image Understanding" (MMIU)
A workshop on indexing images for search.
Funchal, Madeira, Portugal, 22 - 25 January, 2008.
http://www.visapp.org/index.htm

The 16th BOBCATSSS Symposium
An LIS conference with a theme on Information Access, of interest to members who work in the field of Information Seeking.
Zadar, Croatia, 28-30 January 2008
http://www.bobcatsss2008.org/

First International Conference on Web Search and Data Mining (WSDM) 2008
Of particular interest to members interested in search and data mining.
Stanford University, Stanford, California, USA, 11-12 February 2008
http://wsdm2008.org/

23rd Annual ACM Symposium on Applied Computing
A general computing conference with several track of interest to members including an information access track.
Ceará, Brazil, 16-20 March 2008
http://www.acm.org/conferences/sac/sac2008/

AAAI Spring Symposium on Social Information Processing
For those members interested in search and categorisation for Web 2.0.
Stanford University, California, USA, March 26-28, 2008
http://www.isi.edu/~lerman/sss07/

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/

Second International Conference on Weblogs and Social Media
For those members interested in search and categorisation for Web 2.0.
Seattle, Washington, U.S.A, March 31-April 2, 2008
http://www.icwsm.org/2008/

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.