The RIAO 2007 Conference – A Personal View

Adrian Popescu
CEA LIST
18 Route du Panorama, 92260 Fontenay aux Roses, France
adrian.popescu@cea.fr

Abstract

This review summarizes the main topics covered by the presentations given at the RIAO 2007 Conference. The theme of this year's conference, "Large-scale Semantic Access to Content (Text, Image, Video and Sound)", encompasses two current, and recurring, challenges in Information Retrieval: inclusion of semantics in retrieval and adaptation to different types of media. The use of the term "large-scale" underlines the need for scalable methods in real-world applications. RIAO 2007, hosting both innovative applications and research papers, presented a panorama of current trends in semantics based information retrieval.

1 Conference History and Structure

RIAO 2007[1] was held at Carnegie Mellon University in Pittsburgh PA, U.S.A, from May 30 to June 1, 2007. This was the eighth session of the tri-annual conference, held alternately in Europe and in North America. Previous meetings took place in: 1985 (Grenoble); 1988 (M.I.T. Cambridge, MA. USA), 1991 (Barcelona, Spain), 1994 (Rockefeller University, New York, N.Y. U.S.A.), 1997 (McGill University, Montreal, Canada) (2000, Collège de France, Paris) and 2004 (University of Avignon, Vaucluse, France). This year's theme: "Large-Scale Semantic Access to Content (Text, Image, Video and Sound)" was covered by almost 90 presentations, in three different tracks:

- full-papers – 11 sessions comprising a total of 33 papers (acceptance rate 25%)
- posters – 2 sessions including 29 presentations (acceptance rate 33%)
- applications – 20 demonstrations of research prototypes and production systems

[1] The conference was organized by the CID (Centre de Hautes Etudes Internationales d'Informatique Documentaire – the International Center for Advanced Studies in Documentary Information) and its American affiliate CASIS (Center for Advanced Studies on Information Systems), in collaboration with IRIT (Institut de Recherche en Informatique de Toulouse). RIAO is a French acronym for *Recherche d'Information Assistée par Ordinateur* (Computer-aided Information Retrieval). The Organizing Committee was coordinated by Ms. Agnès Bériot, the C.I.D general secretary. The Program Committee was co-chaired by David Evans (CEO, Clarvoyance Corporation for the Americas), Sadaoki Furui (Tokyo Institute of Technology for Asia and Oceania) and Chantal Soulé-Dupuy (IRIT – Université Toulouse I for Europe and Africa) and included 80 members. The Applications Committee numbered 29 members and was co-chaired by Frédéric Le Roux (Agence Régionale de Développement) et Josiane Mothe (IRIT, Université de Toulouse, IUFM, France).

The presenters, as well as the organizers and the attendees, reflected the strong international character of the conference: 10 European and 6 Asian countries were represented, along with the USA, Canada, Australia and Tunisia. The majority of presenters came from France, the USA, Japan and the UK.

2 Presentations

Research in Information Retrieval can be categorized along multiple dimensions, focusing, for example, on the technical paradigm, the research field, the targeted document type, or the application domain. In the following review of RIAO, we retained whichever dimension seemed best for grouping the presentations meaningfully. To give a more complete overview of the conference, we discuss separately a series of applications, the invited talks and enumerate some general conclusions about RIAO 2007. Presentations are found at http://riao.free.fr/program.htm

2.1 Statistical-based Presentations

This section reviews some papers that are based on the use of statistical measures. Targeted research fields include text, sound and image retrieval.

- H. Joho (Univ. of Glasgow, UK) and M. Sanderson (Univ. of Sheffield, UK) carried out a large scale experiment investigating the relation between document frequency and term specificity. The Google corpora and the TREC newspapers set were used to assess the above relation. Terms with different generality degrees (extracted from WordNet) were evaluated and the main reported finding is that document frequency is a more accurate measure for term specificity in the case of the most specific terms in WordNet than in the case of general ones.
- J. Peng (University of Glasgow, UK) et al. introduced a probabilistic model combining document priors (query-independent features) to improve the retrieval effectiveness. Features like the number of a document's incoming links, page-rank, information to noise ratio and document URL type are conditionally combined. Their method is shown to enhance retrieval quality over two large-scale standard collections (GOV and GOV2).
- J. Lu and J. Callan (Carnegie Mellon University, USA) assessed a content-based method for searching over peer-to-peer networks. The effectiveness of the approach is conditioned by the existence of a network overlay that organizes the content (clusters peers detaining similar content that are localized in small parts of the network) and provides good navigability (small world properties hold for the given network). The algorithm for constructing such an overlay is described in the paper, as well as a detailed evaluation that demonstrates its utility.
- R. Neumayer and A. Rauber (Vienna University of Technology, Austria) presented a method for retrieving music based on multi-modal information (acoustic content and lyrics). Extraction of acoustic features relies on the computation of the Bark-scale spectrogram, a psycho-acoustic representation of the signal, while the text processing uses TF-IDF. These two types of information are structured using self-organizing maps, hyperlinked to provide a multi-modal view to the music.
- Content based access to a different type of sound files, orca vocalizations, was explored by G. Tzanetakis (University of Victoria, Canada) et al. who introduced ORCHIVE. Some results concerning the de-noising and classification of recorded vocalizations based on machine learning techniques (neural networks and SVMs) are described as the initial step towards a content based retrieval process over ORCHIVE items.

- C. Millet (CEA-LIST, France) et al. discussed two methods for automatically finding the color of objects and the advantages of employing this information in image segmentation and Web images re-ranking tasks. Statistics concerning the correlation between an object name and different colors are derived using their representation on the Web. Then, an automatic color finding method based on picture segmentation over image sets corresponding to objects is described. In addition to color, several parameters defining the representativeness of a picture are used (for a segmented area, centrality, size and inclusion in the margins of the photo are added). A significant improvement of retrieval precision is obtained in the re-ranked picture sets.
- Image content is equally analyzed by B. Bai (Rutgers University, USA) et al., in the specialized domain of brain imagery. The authors adapted classical IR methods (TF-IDF, LSI) to content based image retrieval (CBIR) and compared them to methods that encode the information in the entire picture. Results show that the methods imported from text retrieval outperform the classical CBIR measures.
- F. Kang (Michigan State University, USA) et al. presented a paper that investigated a new notion of image similarity, more complex than a simple Euclidian distance (currently used in content based image retrieval). Given a large set of images is available, the likelihood of two pictures being clustered together is considered in the similarity measure. Their experiments show that using this measure improves CBIR precision over an Euclidian distance.
- S. Agarwal and S. Arora (Carnegie Mellon University, USA) introduced a word prediction method for SMS messaging based on the context (composed of the numeric code associated to the word on the keyboard of the phone, the preceding term and its part of speech). Several variations of Markov models and the use of SVMs were evaluated for the task and the results seem to indicate that Markov Models are more appropriate than SVMs. The validity of the obtained results is questionable due to the difference between the evaluation corpus (e-mails) and the short messages typed on mobile phones.
- I. Chibane and B.-L. Doan (SUPELEC, France) discussed a new ranking function for Web documents, accounting both for the presence of the query terms on the page and for the popularity of the page among other pages that respond to the same query. The second term in the function, standing for relevance propagation, is calculated using number of links from other relevant pages to the evaluated one. The approach is tested on two standard corpuses (WT10g and .GOV) and the results show that the newly proposed method outperforms a purely-content based method and the PageRank function. The improvement of the results is counterbalanced by an important drawback: the high complexity of the relevance propagation introduced in the ranking function.

2.2 Linguistics-based Presentations

In this section we summarize some presented works that are based on the use of linguistic techniques. Targeted research fields include: textual querying, information extraction, question answering and automatic translation.

- P.-Y. Berger and J. Savoy (Univ. of Neuchatel, Switzerland) introduced a system that automatically selects the best translation of a query based on a prediction of retrieval performance. The authors compared their approach to a machine translation system and a manual translation and showed that its performances are situated between the two.
- F. Moreau (IRISA, Rennes Univ., France) et al. introduced a text retrieval framework that exploits indexes on three levels of linguistic description (morphological, syntactic and semantic) in order to better describe the semantic content of unstructured documents. Three types of

indexes are created and, as they have different natures, three different IR processes are launched. The results are finally merged using a neural network that assesses the relevance of each query-document pair. The main improvement obtained using the fusion approach over a baseline (stem indexes) concerns the higher stability of the results. While comparable performances are reported for precision and recall, the standard deviation is drastically reduced with the merging method.

- Language modeling using Markov chains for query and document expansion was addressed by G. Cao (University of Montreal, Canada) et al. The authors review existing work and show that usually, only one type of expansion is used. The novelty of their approach consists in a simultaneous expansion of the query and the document, doubled by a multi-step expansion process.
- B. Powley and R. Dale (Macquarie University, Australia) presented a method for accurately extracting citations and author names and the referenced fragments from scientific publications. The main novelty compared to existing work is that the full text of the publications is parsed, instead of analyzing only the references section. The evaluation was performed on a corpus containing 6000 scientific publications and very good results (F-measure higher than 97%) are reported in all three tasks.
- R. Budiu (Palo Alto Research Center, Palo Alto) et al. compared three methods that estimate the semantic similarity between words. Latent Semantic Analysis, Pointwise Mutual Information (PMI) and Generalized Semantic Latent Analysis (GLSA) were evaluated in two tasks: synonymy test and comparison with a word similarity assessed by humans. GLSA behaves best on a small scale test corpus, while PMI outperforms on a larger scale set.
- I. Glöckner (FernUniversität in Hagen, Germany) et al. investigated the logical validation of answers and the merging of answers in multi-stream question answering systems. The logical validation is performed using witnesses (snippets in the target documents sustaining the answers). Answers from several results producers (multiple streams) are considered and merged. The reported performances show that the newly introduced approach outperforms all individual QA systems.
- S. Fissaha Adafre and M. de Rijke (University of Amsterdam, The Netherlands) approached the automatic selection of key sentences in documents. Their research is motivated by the fact that an important part of Web queries are informational and undirected and, in these cases, the retrieval process is enhanced if the important sentences are detected. The main contribution of the paper consists in the combination of a graph-based ranking method and a corpus based evaluation of sentence importance. The approach is validated through experiments that prove the effectiveness of the combination of the two methods compared to a separate use and to other baselines.
- J. Jagarlamudi (Microsoft Research, India) et al. presented an approach to multi document summarization that incorporates an evaluation of the importance of the sentences to be included in the summary. The authors model a real-world situation as the summaries they propose depend on real queries. Fair results improvements over existing systems are reported in the paper.

2.3 Structured Information Presentations

In this section we summarize some works employing structured documents as well as papers discussing the utility of organized knowledge in IR architectures. Targeted research fields include: document corpora structuring, textual querying, automatic ontology learning or image retrieval.

- The growing quantity of XML documents available requires adapted retrieval schemes, exploiting the structure of these documents. G. Wisniewski and P. Gallinari (LIP6, France)

presented a method for transforming unorganized Web documents into semi-structured documents. A perceptron like re-ranking algorithm is used to propose the most relevant documents first. This work is useful as it proposes a semi-automatic method for structured resource creation.

- Y. Mass (IBM, Haifa, Israel) et al. described a method for enriching XML Fragments with database operators, obtaining a more structured query language, while preserving the naturalness of the retrieval process over XML Fragments. The validity of the approach is sustained only with anecdotal examples. More formal experiments should be carried in order to support the claims in the paper.
- K. Pinel-Sauvagnat and M. Boughanem (IRIT, Univ. de Toulouse, France) revisited some existing XML retrieval methods and carry experiments concerning term weighting, component length, contextual relevance and use of structural hints. The obtained results sometimes challenge previously published results and give ideas for future improvements of XML retrieval.

A higher level of information organization is represented by formal ontologies and this type of knowledge bases was treated in a number of presentations at RIAO 07.

- Query expansion using topic models extracted from a manually built ontology is explored by X. Wei and W. B. Croft (University of Massachusetts, USA). The authors exploit the conceptual hierarchy constructed for the Open Directory Project so as to reformulate the initial query. This expansion model is assessed against two baselines: a query likelihood model (QLM) and a relevance model (RM) on several subsets of TREC queries. The reported results place the newly introduced method between the QLM and RM and indicate that the expansion model and the RM work well on different types of queries. Further experiments assess model combination, automatic selection of the retrieval model and automatic query categorization and the authors show that some improvement is obtained when using combined models compared to the case when a RM model is employed.
- M. Baziz (IRIT, Univ. de Toulouse, France) et al. introduced an IR method based on query and document expansion using knowledge in an ontology. A comparison between the minimal sub-trees that contain the sets of nodes in the query, and the document is performed to rank the answers. The proposed evaluation shows that, when expanding both the query and the document indexes using knowledge in an ontology, retrieval precision is improved. The subject of this paper is similar to that of G. Cao et al. because they both treat the simultaneous expansion of queries and documents. Important differences arise from the way the expansion process is realized: Baziz et al. mainly rely on ontological knowledge while Cao et al. exploit a language model based on Markov chains.
- V. Challam (Microsoft Co., Redmond, USA) et al. discussed an ontology driven search personalization scheme. Contextual user profiles (reflecting the user's current interest themes) are structured using an ODP ontology and employed to re-rank the results from a non-contextual search engine. Experiments show that the rank of the clicked documents is higher when the search is personalized.
- D. Picca (Univ. of Lausanne, Switzerland) et al. described an unsupervised approach to ontology learning from text. The method is based on the attribution of supersenses (general meanings like *artefact, person, group* etc.) to document terms and their re-ranking in relation to the frequency of the attributed supersenses in specific domains. As a result, domain-related conceptual hierarchies are obtained and the included knowledge is of good quality. An unsolved problem with the approach is the separation between classes and instances which would result in improved ontological relations.

- A. Popescu (CEA LIST, France) introduced an ontology-based framework for Web image retrieval. A multilingual semantic structure was built using parts of WordNet in three languages and its utility in concept disambiguation and search precision improvement was assessed. Larger-scale evaluations need to be carried in order to validate the approach.

2.4 Annotation and browsing-based presentations

A number of the presented papers focused on the annotation and the browsing and navigation of large scale multimedia collections. The importance of user created content description (videos on YouTube, pictures on Flickr, electronic newspapers) is evident.

- In this arena, J. Lanagan and A. F. Smeaton (Dublin City University, Ireland) presented SportsAnno, a prototype system that integrates video summarisation, written reports and user added comments for football matches. The most important novelty in SportsAnno is the collective annotation of the content which results from supporting discussions between the users.
- G. Cabanac (IRIT, Univ. de Toulouse, France) et al. discussed the same topic, collective annotation, but in a more general setting. A detailed analysis of strengths and weaknesses of individual and collective annotation systems is given, along with possible ways to avoid the limitations (notably the unavailability of methods for social validation of annotations) of current applications. The authors describe a prototype implementing their approach, but no evaluation which would validate it in practice is included.
- A related subject was approached by T. Sakai (NewsWatch Inc., Japan) et al., who presented Pic-A-Topic, a prototype system that allows the user to select video segments based on their topic. The system processes TV shows that can be split into stand-alone pieces and possibly be clustered topically. Cue phrase and vocabulary shift detection for topic segmentation are evaluated both separately and using a fusion method and the best results are obtained in the last setting. The overall accuracy of the system is 82% that of a manual segmentation.

2.5 Miscellaneous

This section groups a number of interesting papers that were difficult to be classified elsewhere.

- A. R. Doherty (Dublin City University, Ireland) et al. collected personal data from a camera (SenseCam) and an audio recorder that were worn all day long, as a way of augmenting the human memory. As high volumes of data are recorded, they are easy to exploit only if automatically segmented into meaningful chunks. In all, five information channels were available: images, white light level, acceleration, and temperature from the SenseCam and audio recordings. Experiments have been carried to assess the segmentation results in different settings. The main reported finding is that a combination of the first three enumerated parameters performs better than all other combinations.
- A. Chandramouli and S. Gauch (University of Kansas, USA) argued that a hefty chunk of the Internet traffic is due to the crawling performed by search engines. They discuss an approach meant to reduce this type of traffic via the use of a Web service exploiting Web logs and file system information. At the same time, their crawling method is fitted for finding pages from the hidden Web (dynamically generated content that is not indexed by existing search engines). The evaluation section supports their claims concerning the reduction of the traffic and an improved handling of the hidden Web.

- P. Hanna and P. Ferrero (LaBRI, Univ. de Bordeaux, France) described experiments to optimize local edit algorithms for evaluating melodic similarity. Melody is represented using pitch and note duration. Average Dynamic Recall is calculated in different settings and the optimizations proposed by the authors are useful, as demonstrated by the fact that they obtained the best results in the MIREX 2006 evaluation campaign when assessing symbolic melodic similarity.
- S. Furui (Tokyo Institute of Technology, Japan) presented an overview of existing speech summarization methods, underlining the main steps in the process and the main challenges that appear in this domain. Speech summarization is close to written text summarization, but supplementary difficulties arise as it is necessary to deal with spoken text and this implies a speech recognition process, a non trivial task. The author presents various approaches to speech summarization and concludes that, in spite of an important research effort, the performances of existing systems are considerably worse than manual speech summarization.

2.6 Applications

The applications track illustrated the diversity of problems to be dealt with in information retrieval and we briefly describe here a part of the demonstrations.

- H. Hideki (JustSystems Corporation) demonstrated the *xfy* technology that deals with XML documents, allowing their creation, integration, transformation and management. A unification of arbitrary XML vocabularies is proposed, as well as search functionality in tera-scales databases.
- R. Valdez-Perez (Vivissimo Inc.) presented another search engine dealing with semi-structured data, specifically e-mails.
- J. Pinquier (IRIT, France) introduced ACADI, a system that performs the identification of the main characters in a video (using image and speech recognition) and structures documents using this information.
- H. Do-Duy (SPIKENET Technology) demonstrated MIND, a platform for indexing and retrieving images in videos based on neural networks. One key feature of the application is that the search process is very rapid, taking less than a second for comparing a picture to a database of 30 millions.
- E. Paquet (Institute for Information Technology, National Research Council, Canada) demonstrated Cleopatra, an anthropometric 3D search engine. The fields of application of Cleopatra include: transportation, ergonomics, virtual mannequins and fashion.
- METIOREW is a personalized search application presented by D. Bueno Vallejo (Universidad de Malaga, Spain). A user model is incorporated in the search process and this allows the system to present custom-made results and to give recommendations according to the model.
- Turn Inc. presented a platform for automatic on-line advertising targeting that exploits machine learning techniques. The targeting process is based on the evaluation of parameters like: past performance, brand strength, user profiles or site categories.
- Find Inc. demonstrated their platform for semantic information extraction from Web sites and an application of the method for on-line shopping.
- P. Kislin (Loria, France) introduced METIORE-WISP, a tool that also performs semantic information organization using Web pages as raw data sources. Endeca presented an information access platform that allows a guided search through heterogeneous data.

Still images were the subject of two systems.

- NPPICTURE (New Phenix) incorporates content-based image retrieval techniques and linguistic technologies for analyzing natural language queries and translating them into several languages in order to obtain more image results.
- J. Woods (University of Essex, UK) presented an analysis toolkit based on image regions. The application segments the images, extracts objects and binary partition trees.

2.7 Invited talks

Besides regular presentations, RIAO 2007 included four high-quality invited talks targeting topics such as the challenges related to information fusion and its utility in IR frameworks; the main challenges faced when passing from IR research to the commercial world; the hard choice between following well-established research paths and exploring new ones; and the notion of context, both from a document and an user perspective.

- Jaime Carbonell (Carnegie Mellon University, USA) introduced topics related to information fusion and presented his vision about the information search of tomorrow. In this vision, next-generation search engines deal with multiple dimensions of information: popularity, novelty, trustworthiness and appropriateness to user. The author insists on the need for a semantic description of content as a prerequisite for proposing efficient search applications and grants a central role to this type of applications if one wants to use the Web as a source of knowledge.
- David Evans (Just Systems Evans Research, Inc.) listed a series of challenges coming from the commercial world addressed to researchers in IR. The size and the persistency of data to be handled by IR applications are two critical points in a real-world context. The structural complexity of the data and the solutions to cope with it (formatting, annotations) are discussed next. This problem is closely related to the heterogeneity of the documents, which can have several degrees of formal organization and different natures (text, images, and videos). In each this context, the author introduces the notion of "Mash-up" or Composite Information Objects a form of information fusion that can constitute the basis of a processing across media. Another dimension of heterogeneity is represented by the rights associated to data and this becomes important in a context where huge amount of information comes from a large variety of sources. The semantics of documents is another analyzed point, with a number of facets like: the reference to concepts, their representation and relations, the metrics used to retrieve useful semantic information and the communication of results to the user. Finally, faith is seen as an important factor in a context where there is no common methodology for evaluating the results.
- Donna Harman (NIST, USA) showed that the choice between moving on with existing IR paradigms and shifting to new ones is entirely contextual. Shifting to a new paradigm is definitely a solution when the results of current methods do not improve over a long period and this assertion is supported with an example over a period of seven years from the TREC evaluation campaign where the results for the last three years were quasi-identical. The author underlines the importance of failure analysis when one wants to continue working using the same approach because this type of analysis allows us to spot the main problems with a particular method. The discussion about failure analysis continued during the conference and the organization of a dedicated workshop was considered. The usefulness of such an initiative is indisputable and we hope such a workshop will be held.
- Alan Smeaton (Dublin City University, Ireland) focused on the relation between the content of multimedia documents and the different types of contexts one can associate with these items.

First, the speaker underlined that, in the past, IR research concentrated on the analysis of the content of documents and that today, there is a shift towards an analysis of the context. Second, the notion of context is separated into document-related and user-related context and it is argued that the first type is better explored than the second. Third, A. Smeaton details the notion of user context, divided into: mental, physical and social context. It is rightfully assumed that the user contexts are more difficult to be formalized than document ones as they are highly subjective. Finally, the presentation enumerates some ways of exploiting existing technologies (like Bluetooth or intelligent sensors) in order to obtain contextual information about the user. One possible outcome of such a process is a better targeted advertising.

2.8 Conclusion

The RIAO 2007 conference featured research covering a large array of Information Retrieval related subjects and the presented papers well reflected the state of the art in their respective fields. The overall quality of the papers was very good and they generally reflected the theme of this year's conference: the large scale access to multimedia content. The inclusion of an application track in the conference program facilitated the comprehension of the passage from scientific research to real-world applications.

The non-parallel presentations allowed the attendee to have a fair idea about some of the main challenges faced by the IR community: the necessity to cope with huge (and continuously growing) amounts of data; the necessity of a semantic processing of documents, as a prerequisite for an adaptation of the search process to the users' needs; the adaptation of IR methods to different types of documents; the evaluation issues raised by applications targeting different fields and different.

The proceedings of the RIAO 2007 conference are freely available on the conference site: http://riao.free.fr/.