



**SIXTH FRAMEWORK PROGRAMME
PRIORITY 2
“Information Society Technologies”**



Deliverable D9.3 (Month 12)
SAPIR Publications
December 31, 2007

Project acronym: SAPIR

Project full title: Search on Audio-visual content using Peer-to-peer Information Retrieval

Contract no.: 45128

Deliverable type: Report

Classification: Pub.

Work package and task: WP9, T9.3

Responsibility: ISTI-CNR

Editor: Fausto Rabitti (ISTI-CNR)

Contributors: All partners

Internal Reviewer: Caroline Hagege (XRCE)



EXECUTIVE SUMMARY

This report presents the result of the SAPIR project in the first year, in terms of scientific publications. According to the DOW, the report describes project activities undertaken during the first year of the project, as part of task T9.3. The activity of publishing papers is considered particularly important for the project since it is both an important way of disseminating project results in the research community and a way of receiving valuable feed-backs.

Moreover, for this first report on the project scientific publications, we decided to add a list of papers, published by the project partners before the beginning of the project, which constitute the basis of the project itself.



TABLE OF CONTENTS

EXECUTIVE SUMMARY	2
1 INTRODUCTION	4
2 SCIENTIFIC FOUNDATIONS OF SAPIR.....	5
3 SAPIR SCIENTIFIC PUBLICATIONS – FIRST YEAR.....	9
4 SUMMARY.....	15



1 INTRODUCTION

This report presents the result of the SAPIR project, in the first year, in terms of scientific publications. According to the DOW, the report describes project activities undertaken, during the first year of the project, as part of task T9.3.

Publishing papers on the project results, even if in an study phase or an early development phase, is considered an important way of disseminating project results and receiving feedback on them from the international scientific community. The results are described in Section 3, where all papers are listed, together with a short abstract for each of them.

Moreover, we decided that it is important to show that the leading ideas of SAPIR proposal are based on previous research activities performed by the partners in the different technical areas of the project. Therefore, we present in Section 2 a list of papers published by the partners before the beginning of the project. This research constitutes the basis of the project itself.

2 SCIENTIFIC FOUNDATIONS OF SAPIR

In this Section we give a list of references to papers published by the project partners, before the beginning of the project, and constituting the starting point of the projects in its different areas of research.

IBM papers concern XML querying (WP5, Task 5.1):

- Carmel D. , Maarek Y. ,Mandelbrod M. , Mass Y. & Soffer A.(2003). *Searching XML Documents via XML Fragments*. In the Proceedings of SIGIR' 2003, Toronto, Canada, Aug. 2003.
- Broder, Y. Maarek, Y. Mass, and M. Mandelbrod. *Using XML to Query XML- From Theory to Practice*. In Proceeding of RIAO, 2004

and speech recognition (WP3, Task 3.2):

- J. Mamou, B. Ramabhadran, O. Siohan. *Vocabulary independent spoken term detection*. In Proceedings of SIGIR, 2007.
- Jonathan Mamou, David Carmel, Ron Hoory *Spoken Document Retrieval from Call-Center Conversations*, SIGIR 2006

ISTI-CNR papers concern caching (WP4, Task 4.3):

- T. Fagni, S. Orlando, F. Silvestri, R. Perego. *Boosting the Performance of Web Search Engines: Caching and Prefetching Query Results by Exploiting Historical Usage Data*. ACM Transactions on Information Systems, Vol. 24, n. 1, January 2006.
- T. Fagni, F. Silvestri, R. Perego. *A Highly Scalable Parallel Caching System for Web Search Engine Results*. Proceedings of Euro-Par 2004, Pisa, Italy, August 31-September 3, 2004.
- D. laforenza, C. Lucchese, S. Orlando, R. Perego, D. Puppini, F. Silvestri. *On the Value of Query Logs for Modern Information Retrieval*, chapter in the book Distributed Agent-based Retrieval Tools, A. Soro, G. Paddeu and G. Armano (eds.), Polimetrica International Scientific Publisher, Italy, 2006.

image analysis and retrieval (WP3, Task 3.3):

- G. Amato, F. Falchi, C. Gennaro, F. Rabitti, P. Savino, P. Stanchev. *Improving Image Similarity Search Effectiveness in a Multimedia Content Management System*. In *Proceedings of the 10th Workshop on Multimedia Information Systems (MIS 2004)*, pages 139-146. August 2004.
- F. Falchi, F. Rabitti, W. Schweibenz, J. Simane, *The Web Database of Florentine Coats of Arms. A Project of Identify Complex Image Content Using Both Indexing and Pattern Recognition*, in EVA 2007: Electronic Imaging & the Visual Arts (Firenze, March 26-30, 2007), pp. 187-192.
- F. Falchi, F. Rabitti, W. Schweibenz, J. Simane, *Florentine Coats of Arms on the Web: Experimenting retrieval based on text or image content*, in ISI 2007: 10. Internationales Symposium für Informationswissenschaft (Köln, May 30 - June 1, 2007).

multimedia content management, indexing and retrieval, implemented in the MILOS system (WP5 Task 5.2):

- G. Amato, C. Gennaro, P. Savino, F. Rabitti. *MILOS: A multimedia Content Management System for Multimedia Digital Library Applications*. In First Italian Research Conference on Digital Library Management Systems, Padova, Italy, pages 29-32. January 2005.
- G. Amato, C. Gennaro, P. Savino, F. Rabitti. *Milos: a Multimedia Content Management System for Digital Library Applications*. In Proceedings of the 8th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2004), Volume 3232 of Lecture Notes in Computer Science, pages 14-25. Springer, September 2004.
- G. Amato, C. Gennaro, F. Rabitti, P. Savino. *Milos: A Multimedia Content Management System*. In Proceedings of the 12th Italian Symposium on Advanced Database Systems (SEBD 2004), pages 342-349. LITHOSgrafiche, June 2004.
- G. Amato, P. Bolettieri, F. Debole, F. Falchi, F. Rabitti, P. Savino: *Using MILOS to Build a Multimedia Digital Library Application: The PhotoBook Experience*. ECDL 2006: 379-390



MU-Brno and ISTI-CNR have been cooperating for a long time on similarity based searching in central and distributed environments (WP4, Task 4.1):

- Zezula, Pavel - Amato, Giuseppe - Dohnal, Vlastislav - Batko, Michal. *Similarity Search: The Metric Space Approach*. New York, NY 10013, USA: Springer, 2005. Advances in Database Systems, Vol. 32. ISBN 0-387-29146-6.
- Ciaccia, Paolo - Patella, Marco - Zezula, Pavel. *M-tree: An Efficient Access Method for Similarity Search in Metric Spaces*. In Proceedings of 23rd International Conference on Very Large Data Bases. San Fransisco, California: Morgan Kaufmann, 1997. pp. 426-435. ISBN 1-55860-470-7.
- Dohnal, Vlastislav - Gennaro, Claudio - Savino, Pasquale - Zezula, Pavel. *D-Index: Distance Searching Index for Metric Data Sets*. Multimedia Tools and Applications, Kluwer Academic Publishers, 21, 1, pp. 9-33. ISSN 1380-7501. 2003.
- Batko, Michal - Gennaro, Claudio - Zezula, Pavel. *Similarity Grid for Searching in Metric Spaces*. In Peer-to-Peer, Grid, and Service-Oriented in Digital Library Architectures: 6th Thematic Workshop of the EU Network of Excellence DELOS. Revised Selected Papers. LNCS 3664. Berlin : Springer-Verlag Heidelberg, 2005. pp. 25-44. ISBN 3-540-28711-6.
- Batko, Michal - Gennaro, Claudio - Zezula, Pavel. *A Scalable Nearest Neighbor Search in P2P Systems*. In 2nd International VLDB Workshop on Databases, Information Systems and Peer-to-Peer Computing. Revised Selected Papers. LNCS 3367. Berlin: Springer-Verlag Heidelberg, 2005. pp. 79-92. ISBN 3-540-2523
- Falchi, Fabrizio - Gennaro, Claudio - Zezula, Pavel. *A Content-Addressable Network for Similarity Search in Metric Spaces*. In Databases, Information Systems and Peer-toPeer Computing. Trondheim, Norway : Tapir Uttrykk, 2005. pp. 126-137.
- Batko, Michal - Novák, David - Falchi, Fabrizio - Zezula, Pavel. *On Scalability of the Similarity Search in the World of Peers*. In InfoScale'06: Proceedings of the 1st international conference on Scalable information systems. New York, NY, USA: ACM Press, 2006. 12 pages. ISBN 1-59593-428-6

and more by MU-Brno on the same topic (WP4, Task 4.1):

- Novák, David - Zezula, Pavel. *M-Chord: A Scalable Distributed Similarity Search Structure*. In InfoScale '06: Proceedings of the 1st international conference on Scalable information systems. New York, NY, USA: ACM Press, 2006. 10 pages. ISBN 1-59593-428-6
- Batko, Michal - Dohnal, Vlastislav - Zezula, Pavel. *M-Grid: Similarity Searching in Grids*. In Proceedings of International Workshop on Information Retrieval in Peer-to-Peer Networks, ACM CIKM 2006. Arlington : ACM Press, 2006. pp. 17-24. ISBN 1-59593-531-2

MPI papers concern P2P textual information retrieval, implemented in the MINERVA system (WP4 Task 4.2):

- Matthias Bender, Sebastian Michel, Josiane Xavier Parreira, Tom Crecelius: *P2P Web Search: Make It Light, Make It Fly*. CIDR 2007: 164-168
- Sebastian Michel, Matthias Bender, Nikos Ntarmos, Peter Triantafillou, Gerhard Weikum, Christian Zimmer: *Discovering and exploiting keyword and attribute-value co-occurrences to improve P2P routing indices*. CIKM 2006: 172-181
- Matthias Bender, Sebastian Michel, Peter Triantafillou, Gerhard Weikum, Christian Zimmer: *Improving collection selection with overlap awareness in P2P search engines*. SIGIR 2005: 67-74
- Matthias Bender, Sebastian Michel, Peter Triantafillou, Gerhard Weikum, Christian Zimmer: *MINERVA: Collaborative P2P Search*. VLDB 2005: 1263-1266

distributed top-k querying (WP4 Task 4.2):

- Sebastian Michel, Peter Triantafillou, Gerhard Weikum: *KLEE: A Framework for Distributed Top-k Query Algorithms*. VLDB 2005: 637-648
- Sebastian Michel, Thomas Neumann: *Search for the Best but Expect the Worst - Distributed Top-k Queries over Decreasing Aggregated Scores*. WebDB 2007

distributed link analysis (WP5, Task 5.5):

- Josiane Xavier Parreira, Debora Donato, Sebastian Michel, Gerhard Weikum: *Efficient and Decentralized PageRank Approximation in a Peer-to-Peer Web Search Network*. VLDB 2006: 415-426

and semantic overlay networks (WP7):

- Josiane Xavier Parreira, Sebastian Michel, Matthias Bender, Tom Crecelius, Gerhard Weikum: *P2P Authority Analysis for Social Communities*. VLDB 2007: 1398-1401
- Josiane Xavier Parreira, Sebastian Michel, Gerhard Weikum: *p2pDating: Real life inspired semantic overlay networks for Web search*. Inf. Process. Manage. 43(3): 643-664 (2007)

TELENOR papers concern hyperlink-based approaches (WP5, Task 5.5):

- Johannes Bjelland, Geoffrey S Canright, and Kenth Engø-Monsen, *Web Link Analysis: estimating a document's importance from its context*. Invited review article, Teletronikk, 2007.
- Johannes Bjelland, Mark Burgess, Geoffrey S Canright, and Kenth Engø-Monsen, *Eigenvectors of Directed Graphs and Importance Scores: Dominance, T-Rank, and Sink Remedies*. Data Mining and Knowledge Discovery.

UPD papers concern music retrieval (WP3, Task 3.4):

- Orio N. (2006). *Music Retrieval: A Tutorial and Review*. Foundations and Trends in Information Retrieval. vol. 1(1), pp. 1-90 ISSN: 1554-0669.
- Melucci M., Orio N. *Combining Melody Processing and Information Retrieval Techniques: Methodology, Evaluation, and System Implementation*. Journal of the American Society for Information Science and Technology. (2004) vol. 55(12), pp. 1058-1066 ISSN: 1532-2882.

and several aspects of information retrieval (WP4, Task 4.2 and WP7):

- M. Agosti, L. Pretto. *A Theoretical Study of a Generalized Version of Kleinberg's HITS Algorithm*. Information Retrieval, 2005, 8, 219-243.
- M. Agosti and M. Melucci. *Information Retrieval on the Web*. In Lectures in Information Retrieval, M. Agosti, F. Crestani and G. Pasi, editors, Lectures Notes on Computer Science, Springer-Verlag, 2001.
- Bacchin, M., Ferro, N., and Melucci, M. *A Probabilistic Model for Stemmer Generation*. Information Processing & Management, (2005). 41(1):121-137.
- Giorgio Maria Di Nunzio. *Visualization and Classification of Documents: A New Probabilistic Model to Automated Text Classification*. Bulletin of the IEEE Technical Committee on Digital Libraries (IEEE-TCDL), 2(2), 2006
- Giorgio Maria Di Nunzio and Nicola Ferro. *Scientific evaluation of a DLMS: A service for evaluating information access components*. In Julio Gonzalo, Costantino Thanos, M. Felisa Verdejo, and Rafael C. Carrasco, editors, Research and Advanced Technology for Digital Libraries, 10th European Conference, ECDL 2006, Alicante, Spain, September 17-22, 2006, Proceedings, volume 4172 of Lecture Notes in Computer Science, pages 536-539. Springer, 2006.
- Agosti, M. and Ferro, N. *Managing the Interactions between Handheld Devices, Mobile Applications, and Users*. Chapter X. In Lim, E. P. and Siau, K., editors, Advances in Mobile Commerce Technologies, (2003). pages 204-233. Idea Group, Hershey, USA.

XRCE papers concern natural language processing and text summarization (WP3, Task 3.5):

- Frédéric Roulland, Aaron Kaplan, Stefania Castellani, Claude Roux, Antonietta Grasso, Karin Pettersson, Jacki O'Neill *Query Reformulation and Refinement Using NLP-Based Sentence Clustering* The 29th European Conference on Information Retrieval (ECIR), 2-5 April 2007, Rome, Italy.
- Caroline Brun, Caroline Hagege *Extraction d'information en domaine restreint pour la génération multilingue de résumés ciblés*, TALN 2004, Fez, Morocco, April 2004
- Caroline Brun, Caroline Hagege. *Intertwining deep syntactic processing and named entity detection* In ESTAL 2004, Alicante, Spain, October 2004



- Caroline Brun, Caroline Hagege *Normalization and paraphrasing using symbolic methods*, ACL 2003, Second International workshop on Paraphrasing, Paraphrase Acquisition and Applications, Sapporo, Japan, July 7-12, 2003.

EURIX papers concern audio/video analysis and management, implemented in PrestoSpace (WP3, Task 3.2):

- L. Boch, G. Dimino, A. Messina, W. Bailer, R. Basili, W. Allasia, M. Vigilante, C. Bauer: *The PrestoSpace Metadata Access and Delivery Platform*, Proceedings of AXMEDIS2007 - Industrial application and/or demonstrations workshop, 2007.
- W. Allasia, M. Porzio, M. Vigilante, L. Boch, G. Dimino, A. Messina: *PrestoSpace Publication Platform: A System for Searching and Retrieving Enriched Audiovisual Materials*, Proceedings of 7th Workshop of the Multimedia Metadata Applications at I-MEDIA'07, 2007
- Messina, L. Boch, G. Dimino, W. Bailer, P. Schallauer, W. Allasia, M. Groppo, M. Vigilante, R. Basili: *Creating Rich Metadata in the TV Broadcast Archives Environment: The PrestoSpace Project*, Proceedings of AXMEDIS2006 - Industrial application and/or demonstrations workshop, Los Alamitos, CA, USA, 2006.
- R. Basili, M. Cammisa, L. Boch, A. Messina, G. Dimino, V. Tablan, B. Popov, W. Bailer, W. Allasia, M. Vigilante: *From video segmentation to semantic indexing : the PrestoSpace approach*, ESA-EUSC 2006 workshop on Image Information Mining, Torrejon, Madrid, Spain, 2006.



3 SAPIR SCIENTIFIC PUBLICATIONS – FIRST YEAR

In this Section we present the scientific publications of the first year of the project. A short abstract is associated to each paper. Papers are grouped according to the project topics.

WP3, Task 3.4:

M. Agosti, E. Di Buccio, G.M. Di Nunzio, N. Ferro, M. Melucci, R. Miotto, N. Orio, *Distributed Information Retrieval and Automatic Identification of Music Works in SAPIR*, In: M. Ceci, D. Malerba, L. Tanca (Eds.). Proceedings of the Fifteenth Italian Symposium on Advanced Database Systems, SEBD 2007, 17-20 June 2007, Fasano, BR, Italy. ISBN 978-88-902981-0-3, pp. 479-482

Abstract

In the effort to model theoretically a distributed and heterogeneous environment, as the one where SAPIR project is placed, a key issue is capturing the uncertain nature of the retrieval process and routing mechanisms that characterize a distributed context, like the P2P one. A probabilistic model meets this need. We study explicitly the event space for distributed information retrieval, proposing two different approaches, because the event space definition sets not only the model correctness, but also its capability of describing features of the considered context.

The approach has been applied to the retrieval of music documents. To this end, we proposed a novel methodology for the identification of music works from the recording of a performance, yet independently from the particular performance. The methodology fits with a distributed architecture because of its high interoperability and could be exploited perfectly also in the architecture proposed by the SAPIR project.

A prototype system has been designed and developed in order to test the architecture. Tests have been done figuring out a situation of a peer storing a part of the database composed of 200 recordings, all of them representing tonal Western music. Final results returned the 90% of the analyzed recordings ranked among top 3 positions. These results may show that the identification task becomes feasible also when larger collections of competing audio recordings are used.

WP4, Task 4.1:

Michal Batko, David Novak, Pavel Zezula. *MESSIF: Metric Similarity Search Implementation Framework*. In DELOS Conference 2007, Pisa, 13-14 February 2007. Pisa, Italy : ISBN 2-912335-30-2, pp. 11-23. 2007.

Abstract

The similarity search has become a fundamental computational task in many applications. One of the mathematical models of the similarity – the metric space – has drawn attention of many researchers resulting in several sophisticated metric-indexing techniques. An important part of a research in this area is typically a prototype implementation and subsequent experimental evaluation of the proposed data structure. This paper describes an implementation framework called MESSIF that eases the task of building such prototypes. It provides a number of modules from basic storage management to automatic collecting of performance statistics. Due to its open and modular design it is also easy to implement additional modules if necessary. The MESSIF also offers several ready-to-use generic clients that allow to control and test the index structures and also measure its performance.

WP4, Task 4.1:

Fabrizio Falchi, Claudio Gennaro, Fausto Rabitti, Pavel Zezula, *A Distributed Incremental Nearest Neighbor Algorithm*, Infoscale 2007 - The Second International Conference on Scalable Information Systems, Suzhou, China, June 2007.

Abstract



Searching for non-text data (e.g., images) is mostly done by means of metadata annotations or by extracting the text close to the data. However, supporting real content-based audio-visual search, based on similarity search on features, is significantly more expensive than searching for text. Moreover, the search exhibits linear scalability with respect to the data set size. In this paper, we present a Distributed Incremental Nearest Neighbor algorithm (DINN) for finding nearest neighbor in an incremental fashion over data distributed between nodes which are able to perform a local Incremental Nearest Neighbor (local-INN). We prove that our algorithm is optimal with respect to both number of involved nodes and number of local-INN invocations. An implementation of our DINN algorithm, on a real P2P system called MCAN, was used for conducting an extensive experimental evaluation on a real-life dataset.

WP4, Task 4.3

C. Lucchese, S. Orlando, R. Perego, F. Silvestri. *Mining Query Logs to Optimize Index Partitioning in Parallel Web Search Engines*, Infoscale 2007 - The Second International Conference on Scalable Information Systems, Suzhou, China, June 2007.

Abstract

Large-scale Parallel Web Search Engines (WSEs) need to adopt a strategy for partitioning the inverted index among a set of parallel server nodes. In this paper we are interested in devising an effective term-partitioning strategy, according to which the global vocabulary of terms and the associated inverted lists are split into disjoint subsets, and assigned to distinct servers. Due to the workload imbalance caused by the skewed distribution of terms in user queries, finding an effective partitioning strategy is considered a very complex task.

In this paper we first formally introduce Term Partitioning as a new optimization problem. Then we show how the knowledge mined from past WSE query logs can be profitably used to discover good solutions of this problem. Finally, we report many results to show that we are able to effectively reduce both the average number of servers activated per each query, along with the workload imbalance. Experiments are conducted on large query logs of real WSEs.

WP4, Task 4.3:

C. Lucchese, S. Orlando, R. Perego. *Parallel Mining of Frequent Closed Patterns: Harnessing Modern Computer Architectures*, Proceedings of the IEEE International Conference on Data Mining (ICDM), Omaha NE, USA, Oct. 2007.

Abstract

Inspired by emerging multi-core computer architectures, in this paper we present MT_CLOSED, a multi-threaded algorithm for frequent closed item-set mining (FCIM). To the best of our knowledge, this is the first FCIM parallel algorithm proposed so far.

We studied how different duplicate checking techniques, typical of FCIM algorithms, may affect this parallelization. We showed that only one of them allows to decompose the global FCIM problem into independent tasks that can be executed in any order, and thus in parallel.

Finally we show how MT_CLOSED efficiently harnesses modern CPUs. We designed and tested several parallelization paradigms by investigating static/dynamic decomposition and scheduling of tasks, thus showing its scalability w.r.t. to the number of CPUs. We analyzed the cache friendliness of the algorithm. Finally, we provided additional speed-up by introducing SIMD extensions.

WP4 and WP7:

Jan Sedmidubsky, Stanislav Barton, Vlastislav Dohnal, Pavel Zezula. *Querying Similarity in Metric Social Networks*, NBIS 2007 - 1st International Conference on Network-Based Information Systems, Regensburg, Germany, Sept. 2007.

Abstract

In this paper we tackle the issues of exploiting the concepts of social networking in processing similarity queries in the environment of a P2P network. The processed similarity queries are laying the base on which the relationships among peers are created. Consequently, the communities encompassing similar data emerge in the network. The architecture of the presented metric social network is formally defined using the acquaintance and friendship relations. Two version of the navigation algorithm are presented and thoroughly experimentally evaluated. Finally, learning ability of the metric social network is presented and discussed.

WP4 and WP7:

Matthias Bender, Tom Crecelius, Mouna Kacimi, Sebastian Michel, Josiane Xavier Parreira, Gerhard Weikum. *Peer-to-Peer Information Search: Semantic, Social, or Spiritual?* IEEE Data Engineering Bulletin. 30(2): 51-60 (2007).

Abstract

We consider the network structure and query processing capabilities of social communities like bookmarks and photo sharing communities such as del.icio.us or flickr. A common feature of all these networks is that the content is generated by the users and that users create social links with other users. The evolving network naturally resembles a peer-to-peer system, where the peers correspond to users. We consider the problem of query routing in such a peer-to-peer setting where peers are collaborating to form a distributed search engine. We have identified three query routing paradigms: semantic routing based on query-to-content similarities, social routing based on friendship links within the community, and spiritual routing based on user-to-user similarities such as shared interests or similar behavior. We discuss how these techniques can be integrated into an existing peer-to-peer search engine and present a performance study on search-result quality using real-world data obtained from the social bookmark community del.icio.us.

WP5, Task 5.1:

Jonathan Mamou, Yosi Mass, Michael Shmueli-Sheuer and Benjamin Sznajder, *Query Language for Multimedia Content* SIGIR Workshop on New Challenges in Audio-Visual Search, held at 30th SIGIR Conference, July 2007, Amsterdam.

Abstract

The growing amount of digital multimedia data available today and the de-facto MPEG-7 standard for multimedia content description has lead to the requirement of a query language for multimedia content. MPEG-7 is expressed in XML and it defines descriptors of the multimedia content such as audio-visual descriptors, location and time attributes as well as other metadata such as media author, media Uri and more. While most search solutions for multimedia today are based on text annotations, having the MPEG-7 standard opens an opportunity for real multimedia content based retrieval. In this paper we propose an IR-style query language for such multimedia content based retrieval that exploits the XML representation of MPEG-7. The query language is an extension of the "XML Fragments" query language that was originally designed as a Query-By-Example for text-only XML collections. We mainly focus on the unique characteristics of Multimedia content which needs to support similarity search query (range search and K-nearest neighbors) and queries on spatio-temporal attributes.

WP5, Task 5.3 and WP4, Task 4.2:

Claudio Gennaro, Matteo Mordacchini, Salvatore Orlando, Fausto Rabitti *MRoute: A Peer-to-Peer Routing Index for Similarity Searches in Metric Spaces*, 5th International Workshop on Databases, Information Systems and Peer-to-Peer Computing (DBISP2P 2007), held at 33rd VLDB International Conference, Vienna, Austria, Sept. 2007.

Abstract

Similarity search for content-based retrieval (where content can be any combination of text, image, audio/video, etc.) has gained importance in recent years, also because of the advantage of ranking the retrieved results according to their proximity to a query. However, to use similarity search in real world applications, we need to tackle the problem of huge volumes of such mixed multimedia data (e.g., coming from Web sites) and the problem of their distribution on multiple co-operating nodes. This is the situation of the Networked Peers for Business, where the distributed nodes (i.e., peers) represent aggregations of SME's with similar activities and the multimedia objects are descriptions/presentations of their products/services extracted from the companies' Web sites. In this paper we approach this problem by considering a scenario of a network of autonomous peers maintaining a local collection of metric objects (i.e., mixed mode multimedia content). This network forms a distributed Peer-to-Peer search engine for similarity search based on the paradigm of Routing Index. Each peer in the network thus maintains both an index of its local resources and a table for every neighbor, summarizing the objects that are reachable from it. The paper presents techniques that aim to make our P2P similarity-based search system viable, trading approximate results for scalable solutions. Results of simulations that use real collections of images are discussed.

WP5, Task 5.3:

David Novak, Pavel Zezula *LOBS: Load Balancing for Similarity Peer-to-Peer Structures, 5th International Workshop on Databases, Information Systems and Peer-to-Peer Computing (DBISP2P 2007), held at 33rd VLDB International Conference, Vienna, Austria, Sept. 2007.*

Abstract

The concept of peer-to-peer structures has recently been applied on the problem of large-scale similarity search. This resulted in systems where the computational load of the peers is of a high importance. Since no current load-balancing technique is designed for structures of this kind, we propose LOBS -- a general system for load-balancing in peer-to-peer structures with time-consuming searching. LOBS is based on the following principles: measuring the computational load, separation of the logical and the physical level of the system, and detailed analysis of the load source to exploit either data relocation or replication.

This work contains results of experiments we conducted using a prototype implementation of LOBS. In these trials, we used a real-life dataset and we varied the number of peers and the distribution of the query traffic in the system. The results show that LOBS is able to cope with any query-distribution and that it improves both the utilization of resources and the performance of the query processing. The costs of balancing are reasonable and are very small if there is time to adapt to a query-traffic. The behaviour of LOBS is independent of the network size.

WP5, Task 5.5:

Josiane Xavier Parreira, Carlos Castillo, Debora Donato, Sebastian Michel, Gerhard Weikum. *The Juxtaposed approximate PageRank method for robust PageRank approximation in a peer-to-peer web search network. To appear in VLDB Journal, 2008.*

Abstract

Link analysis on Web graphs and social networks form the foundation for authority assessment, search result ranking, and other forms of Web and graph mining. The PageRank (PR) method is the most widely known member of this family. All link analysis methods perform Eigenvector computations on a potentially huge matrix that is derived from the underlying graph, and the large size of the data makes this computation very expensive. Various techniques have been proposed for speeding up these analyses by partitioning the graph into disjoint pieces and distributing the partitions among multiple computers. However, all these methods require a priori knowledge of the entire graph and careful planning of the partitioning. This paper presents the JXP algorithm for computing PR-style authority scores of Web pages that are arbitrarily distributed over many sites of a peer-to-peer (P2P) network. Peers are assumed to compile their own data collections, for example, by performing focused Web crawls according to their interest profiles. This way, the Web graph fragments that reside at different peers may overlap and, a priori, peers do not know the relationships between different fragments.

The JXP algorithm runs at every peer, and it works by combining locally computed authority scores with information obtained from other peers by means of random meetings among the peers in the network. The computation on the combined input of two peers is based on a Markov-chain state-lumping technique, and can be viewed as an iterative approximation of global authority scores. JXP scales with the number of peers in the network. The computations at each peer are carried out on small graph fragments only, and the storage and memory demands per peer are in the order of the size of the peer's locally hosted data. It is proven that the JXP scores converge to the true PR scores that one would obtain by a centralized PR computation on the global graph. The paper also discusses the issue of misbehaving peers that attempt to distort the global authority values by providing manipulated data in the peer meetings. An extended version of JXP, coined TrustJXP, provides a variety of countermeasures, based on statistical techniques, for detecting suspicious behavior and combining JXP rankings with reputation-based scores.

WP5, Task 5.6:

Giorgio Maria Di Nunzio *Using Scatterplots to Improve Naive Bayes Text Categorization and Retrieval*, SIGIR Workshop on Information Retrieval Graphical Models., held at 30th SIGIR Conference, July 2007, Amsterdam.

Abstract

The approach presented in the paper introduces the two-dimensional representation of documents which allows documents to be represented on a two-dimensional Cartesian plane which has proved to be a valid visualization tool for Automated Text Classification for understanding the relationships between categories of textual documents, and to help users to visually audit the classifier and identify suspicious training data. In order to obtain the two coordinates in the case of the Naive Bayes classifier, a reformulation of the equation for the decision of classification has to be written in such a way that each coordinate of a document is the sum of two addends: a variable component $P(d | c_i)$, and a constant component $P(c_i)$. When plotted on the Cartesian plane according to this formulation, the documents that are constantly shifted along the x-axis and the y-axis can be seen. This effect of shifting is more or less evident according to which NB model, Bernoulli or multinomial, is chosen. The same reformulation has been applied in the case of the Binary Independence Retrieval model for Information Retrieval with encouraging results.

WP5, Task 5.6:

Emanuele Di Buccio and Massimo Melucci. *Utilizing Event Spaces for Distributed Information Retrieval*, Proceedings of the International Conference on the Theory of Information Retrieval (ICTIR), Budapest, Hungary, 2007.

Abstract

In this paper, a probabilistic approach to modeling distributed Information Retrieval centered around the notion of event space is illustrated. This notion is the underlying issue of all probabilistic models and its structure cannot be ignored when a probabilistic model is being constructed. The importance of defining the event space is not only related to the correctness of the model, but also to describing different architectures. Three different spaces are proposed in this paper for modeling distributed IR. Each space captures different aspects and dictates a distinct function for ranking by probability of relevance.

WP6, Task 6.2:

Walter Allasia, Fabrizio Falchi, Francesco Gallo, Nicola Orio *A Digital Rights Aware Similarity Measure for Multimedia Documents*, Proceedings of MS '07: Workshop on multimedia information retrieval on The many faces of multimedia semantics, Held at ACM Multimedia 2007, Augsburg, Germany, Sept. 2007, ACM Press, pp. 73-80.

Abstract



This paper presents a novel approach to the retrieval of multimedia documents that considers Intellectual Property Rights (IPR) metadata as a multidimensional feature in a metric space. The approach allows us to perform similarity searches on IPR attributes between digital items and to integrate these searches in a common query by example paradigm. The aim of this work is the management of the metadata related to the IPR, both in centralized systems and in networks with indexing capabilities, for text and similarity searches, providing the basic infrastructure enabling the private use and the commercial exploitation as well. Content based similarity search can help both the user to deal with a huge amount of similar items with different licenses and the content providers to detect fake copies or illegal uses, as discussed for the case of images and music.

WP6, Task 6.3:

Walter Allasia, Filippo Chiariglione, Fabrizio Falchi, Francesco Gallo *An Innovative Approach for Indexing and Searching Digital Rights*, AXMEDIS 2007 - 3rd International Conference on Automated Production of Cross Media Content for Multi-channel Distribution, Barcelona, Spain, November 2007.

Abstract

The aim of this work is a proposal for a new approach concerning the management of the metadata related to the Digital Rights in centralized systems or networks with indexing capabilities for both text and similarity searches, providing the basic infrastructure enabling the private use and the commercial exploitation as well. We present an innovative approach that treats the right management metadata as metric objects, enabling similarity search on IPR attributes between digital items. Moreover we show how the content base similarity search can help both the user to deal with a huge amount of similar items with different licenses and the content providers to detect fake copies or illegal uses.

WP7:

Stanislav Barton, Vlastislav Dohnal, Jan Sedmidubsky, Pavel Zezula *Gauging the Evolution of Metric Social Network*, DBISP2P 2007 - 5th International Workshop on Databases, Information Systems and Peer-to-Peer Computing, held at 33rd VLDB International Conference, Vienna, Austria, Sept. 2007.

Abstract

In this paper, we tackle the issues of analyzing the structural evolution of the metric social network. The metric social network operates in a P2P environment where peers maintain their own data and the relationships among them are formed on the basis of the processed similarity queries. The evolution is analyzed by traditional social networking tools -- the characteristic path length and the clustering coefficient. Nonetheless, due to the special structure of the metric social network, own designed gauges -- the average overlap and robustness of description coefficients -- are presented to analyze the structure of emerging communities encompassing similar data.



4 SUMMARY

This report presents the scientific publications of the SAPIR project in the first year. This is part of task T9.3. In this report, we also added a list of papers, published by the project partners before the beginning of the project and constituting the basis of the project itself.

The results shown in this report demonstrate the attention given to the research activities along the directions of the main research challenges of the project (media specific analysis and retrieval, multimedia indexing, P2P query processing, semantic overlay networks, digital rights in P2P, etc.). The resulting papers have been mainly addressed to conferences and workshops related to the specific topics of the project activities. In this way, we could disseminate and discuss our results in audiences with the right specific competence.

The overall scientific results are quite satisfactory for us, not only in terms of number of publications, but also in terms of their spread that covers the most important topics of the project. Another important remark is that several papers have been coauthored by people from different partners in the project. This fact demonstrates the partner's cooperation in the research activities performed in the project.