



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



Deliverable D9.3 (Month 24)
SAPIR Publications, 2nd report
December 31, 2008

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 second year, in terms of scientific publications. According to the DOW, the report describes project activities undertaken during the second 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 feedback. This year, we have added information on the impact of the publications according to the ISI Citation Index (only for journal publications) as suggested by the reviewers. Finally, we will briefly discuss the scientific production of the second year of the project in terms of quality of target journals, conferences and workshops, and in terms of the coverage of the various topics addressed in the project.



TABLE OF CONTENTS

| | |
|--|-----------|
| EXECUTIVE SUMMARY | 2 |
| 1 INTRODUCTION | 4 |
| 2 SAPIR SCIENTIFIC PUBLICATIONS – SECOND YEAR | 5 |
| 3 CONCLUDING REMARKS | 17 |



1 INTRODUCTION

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

Publishing papers on the project results, 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 2, where all papers are listed, together with a short abstract for each of them. In Section 3, we present concluding remarks, with some considerations of the scientific production in the second year of the project.



2 SAPIR SCIENTIFIC PUBLICATIONS – SECOND YEAR

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

This year, as suggested by the reviewers, we have tried to add the impact of the publications. For this purpose, we used a professional tool, that is, the ISI Web of Knowledge of Thomson Reuters. However, this was only possible for journal publications, since only the Journal Citation Index was available.

WP3, Task 3.2:

Jonathan Mamou and Bhuvana Ramabhadran, *Phonetic Query Expansion for Spoken Document Retrieval*, Proceedings of Interspeech 2008, Brisbane, Australia, 22-26 September 2008.

Abstract

We are interested in retrieving information from speech data using phonetic search. We show improvement by expanding the query phonetically using a joint maximum entropy N-gram model. The value of this approach is demonstrated on Broadcast News data from NIST 2006 Spoken Term Detection evaluation.

WP3, Task 3.2:

Jonathan Mamou, Yosi Mass, Bhuvana Ramabhadran, Benjamin Sznajder. *Combination of Multiple Speech Transcription Methods for Vocabulary Independent Search*, Proceedings of SIGIR SSCS Workshop 2008, 24 July 2008, Singapore.

Abstract

Today, most systems use large vocabulary continuous speech recognition tools to produce word transcripts which have indexed transcripts and query terms retrieved from the index. However, query terms that are not part of the recognizer's vocabulary cannot be retrieved, thereby affecting the recall of the search. Such terms can be retrieved using phonetic search methods. Phonetic transcripts can be generated by expanding the word transcripts into phones using the baseforms in the dictionary. In addition, advanced systems can provide phonetic transcripts using sub-word based language models. However, these phonetic transcripts suffer from inaccuracy and do not provide a good alternative to word transcripts. We demonstrate how to retrieve information from speech data by presenting a novel approach for vocabulary independent retrieval combining search on transcripts that are produced according to different word and sub-word decoding methods. We present two different algorithms: the first is based on the Threshold Algorithm (TA); the second uses a Boolean retrieval model on inverted indices. The value of this combination is demonstrated on data from NIST 2006 Spoken Term Detection evaluation.

WP3, Task 3.2 and 3.3:

Walter Allasia, Fabrizio Falchi, Francesco Gallo, Mouna Kacimi, Aaron Kaplan, Jonathan Mamou, Yosi Mass, Nicola Orio. *Audio-Visual Content Analysis in P2P Networks: The SAPIR Approach*, Proceedings of the 19th International Workshop on

Database and Expert Systems Applications (DEXA 2008), September 1-5, Turin, Italy, pp. 610-614, ISBN 978-0-7695-3299-8, IEEE Computer Society, 2008.

Abstract

Content based search in audio-visual collections requires media specific analysis for extracting low level features to be efficiently indexed and searched. We present the SAPIR media framework for analyzing digital content and representing the extracted features in a common schema, used to index and search content in a P2P network. The framework contains splitters of compound objects into simple objects to deal with complex media like videos, using image and speech analyzers. We report usage of this framework in the SAPIR demo.

WP3, Task 3.4:

Riccardo Miotto and Nicola Orio. *Towards a Content-based Music Search Engine*, Proceedings of the Italian Research Conference on Digital Library Systems, Padova, Italy, Jan, 2008.

Abstract

Large collections of multimedia documents need to be provided with new tools for easy access and retrieval. In this paper we present a prototype system, available through a Web interface, that can identify an unknown recording of classical music using a database of reference recordings. An important characteristic is that the identification can be carried out also when the query is the recording of a different performance of the work stored in the database, possibly played by different instruments and recorded with background noise.

WP4, Task 4.1:

Benjamin Sznajder, Jonathan Mamou, Yosi Mass, and Michal Shmueli-Scheuer. *Metric inverted - an efficient inverted indexing method for metric spaces*, ECIR'08 - Efficiency Issues in Information Retrieval Workshop, 30th March - 3rd April 2008, Glasgow, Scotland.

Abstract

The increasing amount of digital audio-visual content accessible today calls for scalable solutions for content based search. The state of the art solutions reveal linear scalability in respect to the collection size due to the large number of distance computations needed for comparing low level audio-visual features. As a result, search in large audio-visual collections is limited to associated metadata only done by text retrieval methods that are proven to be scalable.

Search in audio-visual content can be generalized to search in metric spaces by assuming some distance function on low-level features. In this paper we propose a framework for efficient indexing and retrieval of metric spaces by extending classical techniques taken from the textual IR methods such as lexicon, posting lists and boolean constraints- thus enable scalable search in any metric space and in particular in audio-visual content. We show the efficiency and effectiveness of our method by experiments on a large image collection

**WP4, Task 4.2:**

Vlastislav Dohnal, Jan Sedmidubsky, Pavel Zezula, David Novak. *Similarity Searching: Towards Bulk-loading Peer-to-Peer Networks*, Proceedings of the 1st International Workshop on Similarity Search and Applications, April 11-12, 2008, Cancún, Mexico, pp. 87-94, ISBN 978-0-7695-3101-4, Los Alamitos CA, IEEE Computer Society, 2008.

Abstract

Due to the exponential growth of digital data and its complexity, we need a technique which allows us to search such collections efficiently. A suitable solution seems to be based on the peer-to-peer (P2P) network paradigm and the metric-space model of similarity. During the building phase of the distributed structure, the peers often split as new peers join the network. During a peer split, the local data is halved and one half is migrated to the new peer. In this paper, we study the problem of efficient splits of metric data locally organized by an M-tree and we propose a novel algorithm for speeding the splits up. In particular, we focus on the metric-based structured P2P network called the M-Chord. In experimental evaluation, we compare the proposed algorithm with several straightforward solutions on a real network organizing 10 million images. Our algorithm provides a significant performance boost.

WP4, Task 4.2:

David Novak, Michal Batko, Pavel Zezula. *Content Based Image Retrieval on the Web*. Future Internet Symposium (FIS08), Vienna, Austria, 28-30 Sept. 2008

Abstract

We present and demonstrate capabilities of Multi-Feature Indexing Network (MUFIN), <http://mufin.fi.muni.cz/>. To achieve an independence of the similarity abstraction and the ability to process large collections of data, MUFIN is based on two basic paradigms: (1) the metric space model of similarity, and (2) the concept of structured Peer-to-Peer networks. In the demonstration we will show an interactive image retrieval system which indexes 50 million images - it is one or two orders of magnitude more than any other system designed to this purpose can do. Further more, the scalability of the system grows with constant complexity, which is very important for future internet technologies. To demonstrate the extensibility of the metric-based MUFIN approach we will show another prototype application: a face-recognition and retrieval system.

WP4, Task 4.3:

Fabrizio Falchi, Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fausto Rabitti. *A Metric Cache for Similarity Search*, 6th Workshop on Large-Scale Distributed Systems for Information Retrieval, Napa Valley, California October 26-30, 2008.

Abstract

As Content-Based Image Retrieval systems are shifting their target towards the Web-scale dimension, an important issue becomes designing new scalable solutions that can contribute to make content-based search a valid complement to current text-based, Web image searching services. In this paper, we investigate the possibility of caching the results of content-based queries in order to reduce the overall computing load, and increase scalability in searching for metric similarity very large databases of images. Similarity search in metric spaces is highly exible and bases its generality on the properties of the distance function used do assess the proximity of any pair of objects. The concept of cache-hit which is well-defined in traditional applications of caching, becomes weaker for this search paradigm, where in

addition to exact matches, also similar ones have to be possibly looked-up from the cache. We analyze these issues from both theoretical and practical points of view, and we propose two different caching algorithms which return the set of exact results when present in the cache, or an approximate set of results in the case similar queries were previously cached, and some measure of the approximation quality can be guaranteed. The results of very promising tests conducted on a collection of one million high-quality digital photos, show that it is worth pursuing this research direction since the proposed caching techniques can have a significant impact on performance, like caching on text queries has been proven effective for traditional Web search engines.

WP4, Task 4.3:

Fabrizio Falchi, Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fausto Rabitti, *Caching Content-based Queries for Robust and Efficient Image Retrieval*, EDBT-2009: Extending Database Technologies, March 23-26, 2009, Saint-Petersburg, Russia.

Abstract

In order to become an effective complement to traditional Web-scale text-based image retrieval solutions, content-based image retrieval must address scalability and efficiency issues. In this paper we investigate the possibility of caching the answers of content-based image retrieval queries in metric space, with the aim of reducing the average cost of query resolution and boosting the overall system throughput. Our proposal exploits the similarity between the query object and the cache content, and allows the cache to return approximate answers with acceptable quality guarantee even if the query has never been encountered in the past. Moreover, since popular images that are likely to be used as query have several near-duplicate versions, we show that our caching algorithm is robust, and does not suffer of cache pollution problems due to near-duplicate query objects. We report on very promising results conducted on a collection of one million high-quality digital photos.

WP4, Task 4.4:

Michal Batko, Fabrizio Falchi, Claudio Lucchese, David Novak, Raffaele Perego, Fausto Rabitti, Jan Sedmidubky, Pavel Zezula. *Crawling, Indexing, and Similarity Searching Images on the Web*, SEBD-2008: Sixteenth Italian Symposium on Database Systems, Mondello, Palermo, Italy, June 22-25, 2008.

Abstract

In this paper, we report on our experience in building an experimental similarity search system on a test collection of more than 50 million images, to show the possibility to scale Content-based Image Retrieval (CBIR) systems towards the Web size. First, we had to tackle the non-trivial process of image crawling and descriptive feature extraction, performed by using the European EGEE computer GRID, building a test collection, the first of such scale, that will be opened to the research community for experiments and comparisons. Then, we had to develop indexing and searching mechanisms which can scale up to these volumes and answer similarity queries in real-time. The results of our experiments are very encouraging for future applications.

**WP5, Task 5.3:**

Fabrizio Falchi, Claudio Gennaro, Fausto Rabitti, Pavel Zezula, *Distance Browsing in Distributed Multimedia Databases*, Future Generation Computer Systems, Elsevier, 25 (2009) pp. 64-76. (ISI IMPACT FACTOR: 1.095)

Abstract

The state of the art of searching for non-text data (e.g., images) is to use extracted metadata annotations or text, which might be available as a related information. However, supporting real content-based audio-visual search, based on similarity search on features, is significantly more expensive than searching for text. Moreover, such search exhibits linear scalability with respect to the data set size, so parallel query execution is needed. In this paper, we present a Distributed Incremental Nearest Neighbor algorithm (DINN) for finding closest objects in an incremental fashion over data distributed among computer nodes, each able to perform its local Incremental Nearest Neighbor local-INN) algorithm. We prove that our algorithm is optimum with respect to both the number of involved nodes and the 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.

WP5, Task 5.3 :

Fabrizio Falchi, Mouna Kacimi, Yosi Mass, Fausto Rabitti, Pavel Zezula, *SAPIR: Scalable and Distributed Image Searching*, Proceeding of The Second International Conference on Semantic and Digital Media Technologies (SAMT 2007), Genova, Italy, 5-7 December 2007, Proceedings pp. 11-12.

NOTE: not included in the 2007 report

Abstract

In this paper we present a scalable and distributed system for image retrieval based on visual features and annotated text. This system is the core of the SAPIR project. Its architecture makes use of Peer-to-Peer networks to achieve scalability and efficiency allowing the management of huge amount of data. For the presented demo we use 10 million images and accompanying text (tags, comments, etc.) taken from Flickr. Through the web interface it is possible to efficient perform content based similarity search, as well as traditional text search on the metadata annotated by the Flickr community. Fast complex query processing is also possible combining visual features and text. We show that the combination of content-based and text search on a large scale can dramatically improve the capability of a multimedia search system to answer the users needs and that the Peer-to-Peer based architecture can cope with the scalability issues (response time obtained for this demo over 10 million images is always below 500 milliseconds)

WP5, Task 5.4:

David Novak, Michal Batko, Pavel Zezula. *Web-scale System for Image Similarity Search: When the Dreams Are Coming True*, Proceedings of the 6th International Workshop on Content-Based Multimedia Indexing, 18-20th June, 2008, London, UK.

Abstract

The digital images became a commodity which is searched on the Web as ordinarily as the web pages. However, web-scale engines are using only the text-search technology on



images' annotations while the true content-based similarity engines do not seem to be ready for such scales. In this paper, we show a way which opens doors towards a Web-scale image similarity search. We present a very flexible system based on the metric space model and on the peer-to-peer paradigm; it uses structures M-Chord and M-Tree as its fundamental components and measures the image similarity by a combination of MPEG-7 features. The system has been implemented including a graphical interface for online demonstrations and it currently indexes 10 million images downloaded from the Web. We present performance experiments, which focus on the approximate search and the results show that the system provides high quality answers in a fraction of the response times of the precise answers.

WP5, Task 5.4:

Michal Shmueli-Scheuer, Chen Li, Yosi Mass, Haggai Roitman, Ralf Schenkel, Gerhard Weikum. *Best Effort Top-K Query Processing Under Budgetary Constraint*, ICDE 2009: 25th International Conference on Data Engineering, March 29 – April 4, Shanghai, China.

Abstract

We consider a novel problem of top-k query processing under budget constraints. We provide both a framework and a set of algorithms to address this problem. Existing algorithms for top-k processing are budget-oblivious, i.e., they do not take budget constraints into account when making scheduling decisions, but focus on the performance to compute the final topk results. Under budget constraints, these algorithms therefore often return results that are a lot worse than the results that can be achieved with a clever, budget-aware scheduling algorithm. This paper introduces novel algorithms for budget-aware top-k processing that produce results that are significantly better than those of state-of-the-art budget-oblivious solutions.

WP5, Task 5.4:

Michal Batko, Petra Kohoutkova, Pavel Zezula. *Combining Metric Features in Large Collections*, Proceedings of the 1st International Workshop on Similarity Search and Applications, April 11-12, 2008, Cancún, Mexico, pp. 79-86, ISBN 978-0-7695-3101-4, Los Alamitos CA, IEEE Computer Society, 2008.

Abstract

Current information systems are required to process complex digital objects, which are typically characterized by multiple descriptors. Since the values of many descriptors belong to non-sortable domains, they are effectively comparable only by a sort of similarity. Moreover, the scalability is very important in the current digital-explosion age. Therefore, we propose a distributed extension of the well-known threshold algorithm for peer-to-peer paradigm. The technique allows to answer similarity queries that combine multiple similarity measures and due to its peer-to-peer nature it is highly scalable. We also explore possibilities of approximate evaluation strategies, where some relevant results can be lost in favor of increasing the efficiency by order of magnitude. To reveal the strengths and weaknesses of our approach we have experimented with a 1.6 million image database from Flickr comparing the content of the images by five similarity measures from the MPEG-7 standard. To the best of our knowledge, the experience with such a huge real-life dataset is quite unique.

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.* VLDB Journal, Vol. 17, N. 2, March 2008. (ISI IMPACT FACTOR: 3.181)**

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 5.5:

Johannes Bjelland, Geoffrey S Canright, and Kenth Engø-Monsen. *Web Link Analysis: estimating a document's importance from its context.* Teletronikk, Special issue on Network Analysis, edited by Geoffrey Canright and Kent Engø-Monsen, no 1, 2008.

Abstract

This article gives a pedagogical overview of the main ideas around, and mathematical approaches to, Web link analysis. In addition, we present a novel method for link analysis, and briefly discuss its properties.

WP5, Task 5.5:

Ksenia Shevchuk. *Ranking and clustering of search results: Analysis of Similarity graph.* Master Thesis at the Norwegian University of Science and Technology, Trondheim, Norway, May 31, 2008. Co-supervised by Heri Ramampiaro NTNU and Geoffrey Canright and Kenth Engø-Monsen, Telenor R&I.



Abstract

How to provide a fast ranking algorithm, which is not dependent on the structure of the database—as an alternative to Hyper-link-based ranking algorithms like PageRank, which are dependent on the hyper-link structure of the database—is the topic of this project. The most common search today is web search. Searching using the web search engine like Google is writing some query in the search field, and retrieving the set of documents which corresponds to this query. This is text based search, where the search engine retrieves all the documents which include or not (depends on the type of a query) the query words. The method of ranking that uses similarity graph analysis does not need documents in the database to be linked with hyperlinks. This method does not deal with authority; instead, the method sorts documents with respect to their relevance by using the similarity graph, which is built up by calculating similarity values for all pairs of documents in the database. The document is more relevant if it is strongly similar to the large number of the documents of the retrieved part (is the most central node of this part of the documents). This method can be used in all domains—even those (ex., images) for which all standard forms of analysis (text and hyperlink analysis) are impossible. In this work the similarity matrix will be studied.

WP5, Task 5.6:

Emanuele Di Buccio, Nicola Ferro and Massimo Melucci. *Towards a Superimposed Peer Infrastructure for information Access*, SEBD-2008: Sixteenth Italian Symposium on Database Systems, Mondello, Palermo, Italy, June 22-25, 2008.

Abstract

The Peer-To-Peer paradigm is a promising approach for several distributed applications, among which distributed storage systems. SPINA is a software architecture that aims at encompassing indexing and retrieval of unstructured documents stored in a P2P network. This paper describes the current status of the design and the implementation of this software architecture.

WP5, Task 5.6:

Emanuele Di Buccio, Nicola Ferro and Massimo Melucci. *Content-based Information Retrieval in SPINA*, Proceedings of the Italian Research Conference on Digital Library Systems, Padova, Italy, Jan, 2008.

Abstract

The advent of Peer-to-peer (P2P) networks on the scene of the search engines poses new challenges for Distributed Information Retrieval (IR) and Digital Library (DL) Systems. A software architecture called SPINA (Superimposed Peer Infrastructure for iNformation Access) is described in this paper.

WP6, Task 6.2:

Walter Allasia, Filippo Chiariglione, Angelo Difino, Francesco Gallo, Marco Milanese, Rossano Schifanella. *Digital Rights Metadata Management and Retrieval on Structured Overlay Networks*, Proceedings of the 9th International Workshop on Image Analysis for Multimedia Interactive Services, May 7-9, 2008, Klagenfurt, Austria, pp. 130-133, IEEE Computer Society, 2008.



Abstract

This paper introduces a suitable way for indexing multimedia metadata on a structured Peer-to-Peer overlay network, with special care to the management of rights metadata expressed by MPEG-21. We have selected a suitable subset of MPEG-21 Rights Expression Language elements to be indexed, in order to map governed contents into a flat space and allow insertion and retrieval of digital contents. Furthermore, we present a distributed application built on a structured overlay network enabling the search of multimedia items using rights related information. Our solution is completely decentralized and can be exploited in any MPEG-21 compliant metadata representation.

WP6, Task 6.2:

Walter Allasia, Francesco Gallo, Marco Milanesio, Rossano Schifanella. *Governed Content Distribution on DHT Based Networks*, Proceedings of the 3rd International Conference on Internet and Web Applications and Services, June 8-13, 2008, Athens, Greece, pp. 391-396, IEEE Computer Society, 2008.

Abstract

Peer-to-Peer (P2P) systems are widely used for sharing digital items without structured metadata and in absence of any kind of digital rights management applied to the distributed contents. In this paper we propose the implementation of a prototype application that makes use of a structured P2P system enabling the indexing of complex metadata, used to express digital rights. In this way the media contents are exchanged and played according to the expressed grants. The creation and the consumption of the shared contents can be performed through any MPEG-21 REL compliant software and the application allows indexing and search for both governed and ungoverned contents. The information about the license can be included in the queries and the P2P network can be used to share governed contents (both free and with fee) in a legitimate way. In particular the proposed approach represents a suitable solution for indexing and querying rights complex structures on DHT based networks.

WP7, Task 7.1:

Matthias Bender, Tom Crecelius, Mouna Kacimi, Sebastian Michel, Thomas Neumann, Josiane Xavier Parreira, Ralf Schenkel, Gerhard Weikum. *Exploiting social relations for query expansion and result ranking*. ICDE 2008: 24th International Conference on Data Engineering, April 7-12, 2008, Cancún, México, Proceedings pp. 501-506, 2008.

Abstract

Online communities have recently become a popular tool for publishing and searching content, as well as for finding and connecting to other users that share common interests. The content is typically user-generated and includes, for example, personal blogs, bookmarks, and digital photos. A particularly intriguing type of content is user-generated annotations (tags) for content items, as these concise string descriptions allow for reasoning about the interests of the user who created the content, but also about the user who generated the annotations. This paper presents a framework to cast the different entities of such networks into a unified graph model representing the mutual relationships of users, content, and tags. It derives scoring functions for each of the entities and relations. We have performed an experimental evaluation on two real-world datasets (crawled from deli.cio.us and Flickr) where manual user assessments of the query result quality show that our unified graph framework delivers high-quality results on social networks.

**WP7, Task 7.2:**

Jan Sedmidubsky, Stanislav Barton, Vlastislav Dohnal, Pavel Zezula. *Adaptive Approximate Similarity Searching through Metric Social Networks*, Proceedings of the 24th International Conference on Data Engineering, April 11-12, 2008, Cancún, Mexico, pp. 1424-1426, ISBN 978-1-4244-1837-4, Los Alamitos CA, IEEE Computer Society, 2008.

Abstract

Exploiting the concepts of social networking represents a novel approach to the approximate similarity query processing. We present an unstructured and dynamic P2P environment in which a metric social network is built. Social communities of peers giving similar results to specific queries are established and such ties are exploited for answering future queries. Based on the universal law of generalization, a new query forwarding algorithm is introduced and evaluated. The same principle is used to manage query histories of individual peers with the possibility to tune the tradeoff between the extent of the history and the level of the query-answer approximation. All proposed algorithms are tested on real data and medium-sized P2P networks consisting of tens of computers.

WP7, Task 7.2:

Ralf Schenkel, Tom Crecelius, Mouna Kacimi, Sebastian Michel, Thomas Neumann, Josiane Xavier Parreira, Gerhard Weikum. *Efficient top-k querying over social-tagging networks*. The 31st Annual International ACM SIGIR Conference, 20-24 July 2008, Singapore. Proceedings pp. 523-530.

Abstract

Online communities have become popular for publishing and searching content, as well as for finding and connecting to other users. User-generated content includes, for example, personal blogs, bookmarks, and digital photos. These items can be annotated and rated by different users, and these social tags and derived user-specific scores can be leveraged for searching relevant content and discovering subjectively interesting items. Moreover, the relationships among users can also be taken into consideration for ranking search results, the intuition being that you trust the recommendations of your close friends more than those of your casual acquaintances. Queries for tag or keyword combinations that compute and rank the top-k results thus face a large variety of options that complicate the query processing and pose efficiency challenges. This paper addresses these issues by developing an incremental top-k algorithm with two-dimensional expansions: social expansion considers the strength of relations among users, and semantic expansion considers the relatedness of different tags. It presents a new algorithm, based on principles of threshold algorithms, by folding friends and related tags into the search space in an incremental on-demand manner. The excellent performance of the method is demonstrated by an experimental evaluation on three real-world datasets, crawled from deli.cio.us, Flickr, and LibraryThing.

WP7, Task 7.2:

Tom Crecelius, Mouna Kacimi, Sebastian Michel, Thomas Neumann, Josiane Xavier Parreira, Ralf Schenkel, Gerhard Weikum. *Social recommendations at work*. The 31st Annual International ACM SIGIR Conference, 20-24 July 2008, Singapore. Proceedings pp. 884



Abstract

Online communities have become popular for publishing and searching content, and also for connecting to other users. User-generated content includes, for example, personal blogs, bookmarks, and digital photos. Items can be annotated and rated by different users, and users can connect to others that are usually friends and/or share common interests. We demonstrate a social recommendation system that takes advantages of users connections and tagging behavior to compute recommendations of items in such communities. The advantages can be verified via comparison to a standard IR technique.

WP7, Task 7.2:

Ralf Schenkel, Tom Crecelius, Mouna Kacimi, Thomas Neumann, Josiane Xavier Parreira, Marc Spaniol, Gerhard Weikum. *Social Wisdom for Search and Recommendation*. IEEE Data Engineering Bulletin. 31(2): 40-49 (2008)

Abstract

Social-tagging communities offer great potential for smart recommendation and “socially enhanced” search result ranking. Beyond traditional forms of collaborative recommendation that are based on the item-user matrix of the entire community, a specific opportunity of social communities is to reflect the different degrees of friendships and mutual trust, in addition to the behavioral similarities among users. This paper presents a framework for harnessing such social relations for search and recommendation. The framework is implemented in the SENSE prototype system, and its usefulness is demonstrated in experiments with an excerpt of the librarything community data.

WP7, Task 7.2:

Tom Crecelius, Mouna Kacimi, Sebastian Michel, Thomas Neumann, Josiane X. Parreira, Ralf Schenkel, Gerhard Weikum. *Making SENSE: Socially Enhanced Search and Exploration*. VLDB 2008: International conference on Very Large Data Bases, Auckland, New Zealand, Aug. 23-28, 2008

Abstract

Online communities like Flickr, del.icio.us and YouTube have established themselves as very popular and powerful services for publishing and searching contents, but also for identifying other users who share similar interests. In these communities, data are usually annotated with carefully selected and often semantically meaningful tags, collaboratively chosen by the user who uploaded an item and other users who came across the item. Items like urls or videos are typically retrieved by issuing queries that consist of a set of tags, returning items that have been frequently annotated with these tags. However, users often prefer a more personalized way of searching over such a ‘global’ search, exploiting preferences of and connections between users. The SENSE system presented in this demo supports hybrid personalization along two dimensions: in the social dimension, a search process is focused towards items tagged by users explicitly selected as friends by the querying user, whereas in the spiritual dimension, users that share preferences with the querying user are preferred. Orthogonal to this, the system additionally integrates semantic expansion of query tags to improve search results. SENSE provides an efficient top-k algorithm that dynamically expands the search to related users and tags. It is based on principles of threshold algorithms, folding related users and tags into the search space in an incremental on-demand manner, thus visiting only a small fraction of the social network when evaluating a query. The demonstration uses three



different real-world datasets: a large set of urls from del.icio.us, a large set of pictures from Flickr, and a large set of books from librarything, each together with a large fraction of the corresponding social network of these sites.



3 CONCLUDING REMARKS

This report presents the scientific publications of the SAPIR project in the second year. This is part of task T9.3.

The results shown in this report demonstrate the continuous attention given to the research activities during the progress of the project. These publications have followed the main research directions of the project: media specific analysis and indexing, P2P indexing and query processing, integration of different overlay networks in the processing of queries, use of link analysis, management of digital rights in the P2P infrastructure, exploiting social networks to improve the retrieval process. This year we added information on the impact of the publications according to the ISI Citation Index (only for journal publications) as suggested by the reviewers.

Like last year, most papers have been addressed to conferences and workshops related to the specific topics of the project activities. The reason is that, given the fast publishing track, we could have a fast way of disseminations of results and a fast way to discuss our results in audiences with the right specific competence. It is worth noting that some of the publications have been presented at some of the most important conferences in the database and information retrieval area, for example, VLDB, SIGIR, EDBT, ICDE. In addition, we had two important journal publications (impact factors, according to the ISI Citation Index, are provided) on more mature topics.

The overall scientific results are very satisfactory for us, not only in terms of quality and number of publications, but also in terms of their spread that covers the most important topics of the project. Another important aspect, that demonstrates the level of cooperation in the projects, is that several papers have been coauthored by people from different partners in the project.