SIGIR Workshop Report

The SIGIR Heterogeneous and Distributed Information Retrieval Workshop

Ranieri Baraglia
HPC-Lab
ISTI-CNR, Italy
ranieri.baraglia@isti.cnr.it

Domenico Laforenza
HPC-Lab
ISTI-CNR, Italy
domenico.laforenza@isti.cnr.it

Fabrizio Silvestri
HPC-Lab
ISTI-CNR, Italy
fabrizio.silvestri@isti.cnr.it

1 Introduction

In the last few years there have been the explosion in the use of heterogeneous distributed systems. Ranging from simple Network of Workstations to the more modern and complex Grid systems, the adoption of distributed systems instead of massively parallel supercomputers has been preferred due to their reduced cost of ownership. These kinds of systems pose many challenges in terms of information access, storage and retrieval. Usually, in fact, instead of having collections stored at a single site they are collected, and sometimes managed, at different sites (possibly owned by different institutions). Particular interest has been expressed on architectures and specifications for information retrieval in the context of heterogeneous distributed computing systems.

The workshop have been focused on new methods and algorithms to efficiently and effectively access data distributed over large heterogeneous distributed systems. The workshop particularly encouraged papers addressing the creation and the search in distributed, dynamic information systems as well as papers presenting novel architectural solution for these systems. However, more broadly, papers were solicited on any topic related to information retrieval in distributed architectures, including the topics listed below.

- Meta-IR: coping with heterogeneous distributed data;
- Algorithms for distributed IR;
- Grid IR;
- Peer-to-Peer (P2P) IR;
- Complex queries specification and resolution;
- Collection selection;
- Query routing (especially in Grid and P2P systems);
- Data fusion;
- Query execution scheduling;
- Index structures;
- Specific architectures;
- Scalability;
- Fault tolerance;
- Privacy;
- Applications.

The workshop was held in Salvador, Bahia, Brazil following the SIGIR 2005 conference. Even if the venue of the conference might have teased people to go outside having a swim or taking sun in one of the many beautiful beaches around Salvador, about 20 people attended the workshop. The group was really interested and active throughout the entire workshop. The papers submitted were all of high quality and the people that actually presented them did that in a very clear and precise way. We have only one remarkable point. Despite the fact that in the Call for Paper Grid Computing was mentioned as an interesting topic, none of the major player of this field was neither attracted to submit any paper nor interested in being included into the Program Committee. Even though we took care to contact them directly.

2 Presentations

Here we are going to provide a very brief description of the papers that have been presented during the workshop. The workshop was divided into four sessions. The complete proceedings are available in electronic form at
http://hdir2005.isti.cnr.it/cameraready/cameraready.pdf.
The first session hosted an invited presentation given by Prof. Ricardo Baeza-Yates entitled "Recent Advances on Distributed Information Retrieval". He started with a brief introduction to the attendant on the topic of what modern Information Retrieval is especially in light of the enormous quantity of highly dynamic data available in the Web. Then, he depicted some current trends in distributed (i.e. parallel) information retrieval systems showing some ongoing works on this field. In particular, one interesting thing is a work that has been done in collaboration with Moffat, Webber, and Zobel showing that a pipelined and term-partitioned architecture for a parallel search-engine is superior to the more traditional farm and document-partioned one. At the end of the talk he cited a list of open problems:

- New retrieval models.
- More on Indexing & Searching.
- Quality evaluation (Web, XML).
- Real distributed architecture.
- More on P2P.
- New ranking sources.
- Spamming detection.
- Multimedia.
- Grid computing.

Each of them needs attention especially if thought in light of the current new directions the Web is going to move toward.
On the second session three papers about "Peer-to-Peer" were presented. The first paper entitled "p2pDating: Real Life Inspired Semantic Overlay Networks for Web Search" was presented by Josiane Xavier Parreira. The paper was about a novel method for building an overlay network on the basis of the novel notion of peer-to-peer "dating" - p2pdating. The authors used and tested p2pDating into JXP and MINERVA. JXP is an algorithm for dynamically computing global authority scores, whereas MINERVA is a distributed search

engine implemented by authors in previous works. Both benefit by the use of Semantic Overlay Networks built through p2pdating. Experiments have shown the superiority of the combination of JXP and p2pDating over just JXP. Moreover, p2pDating is also used to compute in a distributed manner the PageRank score of the pages managed by each node. Experiments have shown that this approach is useful too and the real PageRank score may be approximated in a good way.

The second paper entitled "An integrated approach for searching and browsing in heterogeneous peer-to-peer networks" was presented by Henrik Nottelmann. In their paper they proposed a service oriented architecture (SOA) for supporting searching and browsing in a hierarchical network. Their approach is characterized by an high extensibility and reusability. More than proposing new algorithms, their paper proposes a very good architecture exploiting in a highly optimized fashion state-of-the-art technology in both P2P and IR.

The last paper of the session entitled "A Scalable Semantic Indexing Framework for peer-to-peer Information Retrieval" was presented by ChengXiang Zhai. The paper proposes an information retrieval framework that clusters documents similar in semantics into the same group, and stores them on nearby nodes using a two-phase DHT-based indexing method. The proposed system seems to show attractive properties, in particular: (a) tunable tradeoff between search accuracy and efficiency; (b) support for advanced retrieval methods (such as feedback and personalized search). (c) adaptation to the dynamic nature of P2P networks with incremental incorporation of new concepts.

On the third session there were scheduled two papers on "Distributed Systems". Unfortunately the second speaker Mikhail Sogrin was not able to flew in Brazil and withdrew his talk. The only presented paper in this session is the one entitled "Distributed Processing of Conjunctive Queries" presented by Claudine Badue. This paper carefully and extensively discusses several experiments results regarding a parallel Search Engine that exploits a usual inverted file index distributed according to a document partition scheme. The authors assessed a experimental setup for analyzing in detail performance issues of a small distributed WSE processing real queries on a real index. Load imbalance, broker behavior, CPU and I/O costs of local index servers, overall throughput were studied in depth. The figures obtained cannot probably be obtained in a different setup or in a large-scale WSE, but trends and the general behavior should be.

On the last session two presentation were given on "Heterogeneous Systems". The first speaker Jens Graupmann presented a work on "The Light-Weight Semantic Web: Integrating Information Extraction and Information Retrieval for Heterogeneous Environments". In the paper they show a light-weight version of the Semantic Web. By advocating the use of available off-the-shelf Information Extraction tools, they are able to automatically detect and annotate important classes of information that are frequently used in queries, like locations and dates. They eventually propose a query language that is able to exploit the extra annotations and allows novel range and join conditions.

The last speaker of the workshop has been Pierangelo Veltri who presented a quite orthogonal paper entitled "An Architecture for Managing Mass Spectrometry Data for a Distributed

Proteomics Laboratory". Aim of this work has been to present an architecture that may be used to efficiently organize Mass Spectrometer (MS) samples. The presented a system for loading, preprocessing, storing and managing mass spectra data. In the paper they also present a possible distributed architecture for the management of such a kind of data.

3 Conclusions

At the end of the workshop day there have been a discussion about the possible interesting topics in the near future.
All the participants agreed on the fact that due to the very high dynamic nature of the Web in the (not so far) future maybe traditional Search Engines might not be able to fulfill users' need as they were able so far.
One of the major problem, as indicated also in the invited talk of Amit Singhal from Google, is Spam. But not only spam is an issue. There also is the enormous quantity of dynamical data being produce every instant through the Web: blogs, multimedia archives, mailing lists, and so on and so forth, are only examples of what we are talking about. It seems clear that traditional crawling techniques will become soon inadequate. There are some possible interesting solutions to this problem. One might be to let Web server crawl their local collection and to send the indexed local collection to a centralized index. Several problem will arise from this organization. How to trust each sent index? How can we efficiently merge each index? Another possible solution is represented by distributed IR. May be not intended in the traditional sense but also in a hybrid manner. For instance, one could maintain a centralized database for static (or quasi-static) contents, whereas a decentralized index maintained by servers authoritative for certain topics might be designed to index dynamic contents.
Peer-to-Peer may represent another solution to the problem of content freshness. Typical applications of P2P to IR are represented by discovering information dislocated at various sites. Many of the works presented at this workshop considered this problem but seem to not keep into account the problem of the high churn rate characterizing typical P2P networks. Simulations are rarely done in environments where nodes are continuously entering and exiting the network. As noted in previous workshops on P2P and IR, this research area still lacks a good method for modeling large dynamic networks with reasonable hardware resources. Another problem that is felt really important by the P2P IR community is the lacking of good datasets (including relevance judgments) coming from real networks having large query streams. Another issue came up during the discussion. Traditional distributed (or P2P) IR models try to represent a collection by modeling the amount of potential relevant information contained within the collection itself. On the other hand traditional (i.e. centralized) Search Engines developers pose a lot of attention to efficiency. A possible improvement in P2P IR systems usability might be represented by the inclusion within effectiveness-based models of

metrics that would keep into account the efficiency of each collection manager. In this way we could also have systems aware of response times and throughput.
Another interesting question that have been raised during the discussion is the usefulness of a model to predict performance of parallel IR systems. The work presented by Claudine Badue *et al.* goes precisely toward this goal. There have been, though, many people that raised interesting questions. How to model the presence of a cache in front of the system? How to model the behavior of operating system mechanisms (i.e. buffer cache, disk cache, process scheduling, network subsystem, etc.)? How to consider the collection distributed (random or not)?

4 Acknowledgments

We thank all the PC members for their great and precise support in reviewing the papers we received, their comments have certainly been of help for authors in improving their works. We would like to thank also the organizers of the SIGIR conference for giving us the opportunity to set up this workshop and for their continuous support during the months before the workshop took place. Last but not least, we also thanks all the participants at the workshop for their comments, questions and doubts, this workshop has been possible, obviously, because of their work. Thank you all!