CrowdSearch: Crowdsourcing Web search

[A WWW 2012 Workshop] *

Piero Fraternali, Stefano Ceri
Dipartimento di Elettronica e Informazione
Politecnico di Milano, Italy
name.surname@polimi.it

Ricardo Baeza-Yates
Yahoo! Research
Spain
rbaeza@acm.org

Fausto Giunchiglia
University of Trento
Italy
fausto@dit.unitn.it

1. GOALS AMD MOTIVATIONS

Link analysis, that has shaped Web search technology in the last decade, can be seen as a massive mining of crowd-secured reputation associated with pages. With the exponential increase of social engagement, link analysis is now complemented by other kinds of crowd-generated information, such as multimedia content, recommendations, tweets and tags, and each person can ask for information or advices from dedicated sites. With the growth of online presence, we expect questions to be directly routed to informed crowds. At the same time, many kinds of tasks - either directly used for search or indirectly used for enriching content to make it more searchable - are explicitly crowd-sourced, possibly under the format of games. Many such tasks can be used to craft information, e.g. by naming and tagging data objects and by solving representational ambiguities and conflicts, thereby enhancing the scope of searchable objects. Thus, social engagement is empowering and reshaping the search of Web information.

CrowdSearch is targeted to enabling, promoting and understanding individual and social participation to search. It addresses important research questions, such as: How can search paradigms make use of social participation? Will keyword-based search seamlessly adapt to social search, or instead will new models of interaction emerge? Should social interaction be stimulated by curiosity, games, friendship or other incentives? Is there a Şcrowdsearching etiquetteŤ to be used when engaging friend or expert communities? Should new sources of information be socially scouted? Which are the mechanisms that may be used to improve or reshape search results based upon social ranking? How do social ranking models compare to advertising? Will social interaction solve the problems of data integration? What is the role of semantics, and can it help CrowdSearch?

The workshop aims at gathering researchers from different fields to debate about the various concepts, approaches, architectural choices and technical solutions for opening information search to the active participation of human beings. The key idea is that human beings should be actively involved in different stages of the search and their actions should be composed and intermixed with those of computers to get the best possible search results.

CrowdSearch 2012 workshop at WWW 2012, Lyon, France

1.1 Topics of interest

The topics of interest for the CrowdSearch workshop include:

- Large-scale knowledge discovery, content enrichment and quality assessment with the support of humans and communities.
- Models for task crowdsourcing and game creation for information augmentation, integration, extraction, classification, and retrieval.
- Software models, architectures, and tools for combining information management with human and social computations.
- Throughput, processing time, and results quality optimization of queries that involve both data and human sources.
- Incentive mechanisms for engaging users in tasks and games, either individually or cooperatively within social networks.
- Techniques for identifying and mitigating spam and abuse in crowd search tasks.
- Approaches for measuring the effectiveness and quality of human and social applications for information retrieval and their empirical assessment.
- Human and social computation in multimedia content processing for search.
- Use cases and applications of human-assisted information retrieval.
- Role of crowd search in Şbig dataŤ applications.
- User models and human factors in task design for crowdsourced search applications, e.g., cognitive bias, bounded rationality, understanding the boundaries between search questions and spam, etc.

2. PROGRAM HIGHLIGHTS

The Workshop has gathered ten research papers, which have been organized in three Workshop sessions.

2.1 Crowdsearching on textual and linked data

The first session deals with **crowdsearching on textual and linked data**.

The paper by Ali Khodaei and Cyrus Shahabi, *Social-Textual Search and Ranking*, focuses on how to improve the effectiveness of web search by utilizing social data available from users, users actions and their underlying social network, defined as *social-textual* (*socio-textual*) search. They show how social aspects can be effectively integrated into the textual search engines and propose a new social relevance ranking based on several parameters including relationship between users, importance of each user and actions users perform on web documents (objects). The proposed social ranking is combined with the conventional textual relevance ranking and evaluated with experiments based on data the from online radio website *last.fm*.

Elena Simperl, Maribel Acosta and Barry Norton, in the paper *A semantically enabled architecture for crowdsourced Linked Data management*, propose a semantically enabled architecture for crowdsourced data management systems which uses formal representations of tasks and data to automatically design and optimize the operation and outcomes of human computation projects. The architecture is applied to the context of Linked Data management to address specific challenges of Linked Data query processing such as identity resolution and ontological classification. Starting from a motivational scenario they explain how query-processing tasks can be decomposed and translated into MTurk projects using a semantic approach.

In the position paper *Exploiting Twitter as a Social Channel for Human Computation* Ernesto Diaz-Aviles, Ricardo Kawase and Wolfgang Nejdl propose a novel decentralized architecture that exploits the Twitter social network as a communication channel for harnessing human computation. Their framework provides individuals and organizations the necessary infrastructure for human computation, facilitating human task submission, assignment and aggregation. The paper also presents a proof of concept and explores the feasibility of the proposed approach in the light of several use cases.

2.2 Methods and Tools for CrowdSearching

The second session focuses on **Methods and Tools for CrowdSearching**.

In the paper *Human Computation Must Be Reproducible* Praveen Paritosh argues that in the social and behavioral sciences, when using humans as measuring instruments, *reproducibility* guides the design and evaluation of experiments and that the results of human computation, which has similar properties, must be reproducible, in order to be informative. Additionally he discusses the requirements of *validity* or *utility* of results, which depend on reproducibility. Reproducibility has implications for the design of task and instructions, as well as for the communication of the results.

The paper *Mechanical Cheat: Spamming Schemes and Adversarial Techniques on Crowdsourcing Platforms*, by Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux, reviews techniques currently used to detect spammers and malicious workers in crowdsourcing platforms, whether they are bots or humans randomly or semi-randomly completing tasks; then, the authors describe the limitations of existing techniques by proposing approaches that individuals, or groups of individuals, could use to attack a task on existing crowdsourcing platforms. They focus on crowdsourcing relevance judgements for search results as a concrete application of their proposed techniques.

Marco Brambilla, Alessandro Bozzon and Andrea Mauri, in the short paper *A Model-Driven Approach for Crowdsourcing Search*, propose a model-driven approach for the specification of crowd-search tasks. In particular they define two models: the Query Task Model, representing the metamodel of the query that is submitted to the crowd and the associated answers; and the User Interaction Model, which shows how the user can interact with the query model to fulfill her needs. This approach allows for a top-down design, from the crowd-search task design, down to the crowd answering system design, this grants automatic code generation thus leading to quick prototyping of search applications based on human responses collected over social networking or crowdsourcing platforms.

2.3 Crowdsourcing for Multimedia Applications

The third session addresses the specificities of **Crowdsourcing for Multimedia Applications**.

Masataka Goto, Jun Ogata, Kazuyoshi Yoshii, Hiromasa Fujihara, Matthias Mauch and Tomoyasu Nakano, in the paper *PodCastle and Songle: Crowdsourcing-Based Web Services for Retrieval and Browsing of Speech and Music Content*, describe two web services, *PodCastle* and *Songle*, that collect voluntary contributions by anonymous users in order to improve the experiences of users listening to speech and music content available on the web. These services use automatic speech-recognition and music-understanding technologies to provide content analysis results, such as full-text speech transcriptions and music scene descriptions, that let users enjoy content-based multimedia retrieval and active browsing of speech and music signals without relying on metadata. When automatic content analysis is used, however, errors are inevitable. PodCastle and Songle therefore provide an efficient error correction interface that let users easily correct errors by selecting from a list of candidate alternatives.

In the paper *A Framework for Crowdsourced Multimedia Processing and Querying*, Alessandro Bozzon, Ilio Catallo, Eleonora Ciceri, Piero Fraternali, Davide Martinenghi and Marco Tagliasacchi, introduce a conceptual and architectural framework for addressing the design, execution and verification of tasks by a crowd of performers. The proposed framework is substantiated by an ongoing application to a problem of trademark logo detection in video collections. Preliminary results show that the contribution of crowds can improve the recall of state-of-the-art traditional algorithms, with no loss in terms of precision. However, task-to-

executor matching, as expected, has an important influence on the task performance.

Christopher G. Harris, in the paper *An Evaluation of Search Strategies for User-Generated Video Content*, examines user-generated content (UGC) search strategies on YouTube using video requests from several knowledge markets such as Yahoo! Answers. He compares crowdsourcing and student search efforts to YouTubeŠs own search interface, applies these strategies to different types of information needs, ranging from easy to difficult, and evaluatee findings using two different assessment methods and discuss how the relative time and financial costs of these three search strategies affect our results.

Finally, the paper *Discovering User Perceptions of Semantic Similarity in Near-duplicate Multimedia Files* by Raynor Vliegendhart, Martha Larson and Johan Pouwelse, addresses the problem of discovering new notions of user-perceived similarity between near-duplicate multimedia files, with focus on file-sharing, since in this setting, users have a well-developed understanding of the available content, but what constitutes a near-duplicate is nonetheless nontrivial. An experiment elicited judgments of semantic similarity by implementing triadic elicitation as a crowdsourcing task on Amazon Mechanical Turk. The judgments are categorized in 44 different dimensions of semantic similarity perceived by users. These discovered dimensions can be used for clustering items in search result lists.

3. ACKNOWLEDGMENTS

The Crowdsearch workshop is sponsored by the CUBRIK Integrating Project of the 7th Framework Program of the EU. We wish to thank all the members of the Program Committee, who contributed to selecting an attractive program, and the invited speakers Donald Kossman and Sihem Amer-Yahia.