Driving Curiosity in Search with Large-scale Entity Networks

Ilaria Bordino[1], Mounia Lalmas[2], Yelena Mejova[3] and Olivier Van Laere[1]
[1] Yahoo Labs, Barcelona
[2] Yahoo Labs, London
[3] Qatar Computing Research Institute, Doha

In many search scenarios, users sometimes find unexpected, yet interesting and useful results, which make them curious; they experience serendipity. This curiosity encourages them to explore further. We developed an entity search system designed to support such an experience.

The system explores the potential of entities extracted from two of the most popular sources of user-generated content – Wikipedia, a user-curated online encyclopedia, and Yahoo Answers, a more unconstrained question & answering forum – in promoting serendipitous search. The content of each data source is represented as a large network of entities, enriched with metadata about sentiment, writing quality, and topical category. A lazy random walk with restart is implemented to retrieve entities from the networks for a given entity query.

This paper discusses our work, focusing on our experience in designing, developing, and evaluating such a system. We also discuss the challenges in developing large-scale systems that aim to drive curiosity in search.

DOI: 10.1145/2682914.2682919 http://doi.acm.org/10.1145/2682914.2682919

1. MOTIVATION

The classic Web search experience, consisting in returning *ten blue links* in response to a short user query, is powered today by a mature technology. However, the ten blue links represent only a fractional part of the total Web search experience: today, what users expect and receive in response to a web query, is not just *relevant documents*, but a plethora of multi-modal information extracted and synthesized from numerous heterogenous sources on and off the Web. The user search intent has considerably evolved.

Nowadays, web users often enjoy visiting a website without a specific search objective in mind, but rather based on the simple desire to get an update, or be entertained during their spare time. *Searching for fun* or having fun while searching involves activities such as online shopping with nothing to buy, reading online, watching funny videos or finding funny pictures. *Serendipity* (a discovery of something new and interesting) has been an important consideration for recommender systems, and it is becoming increasingly important also for

search systems, which are now constructed, at least partly, with the objective of engaging users, so as to keep them interacting with the system even without a predefined purpose.

In parallel, data complexity and its diversity have been rapidly expanding over the last years, spanning from large amounts of unstructured and semi-structured data to semantically rich available knowledge. Having to face a *web of objects* ([Baeza-Yates 2010]) rather than a web of links, modern search engines have shifted their main goal from relevant document selection towards satisfactory task completion. The richness of data provides search systems with promising opportunities to develop sophisticated discovery capabilities and promote different types of search experience, including *serendipitous search*.

The drive for satisfying increasingly complex user information needs originated efforts around organizing information in semantically meaningful ways, which resulted in creating new standards to annotate the content on the Web with its semantic meaning, or notable resources for organizing knowledge on the Web, such as Wikipedia and DBPedia, which provide powerful means to build multilingual text processing and entity extraction tools.

Other less structured sources of knowledge are collected in different types of social media, which allow the users to interact with each other, exchanging views and opinions. The proliferation of web fora and question-answering web sites such as Yahoo Answers, as well as topic specific ones like Stack Overflow[1] for computer programmers or HealthTap[2] for medical questions, has transformed the Web into a global platform where users are not only consumers, but they actively *produce* content. These less structured resources contain the wisdom – both factual and informal– of thousands of non-professional users, providing timely, community-generated context for topics ranging from celebrities, popular culture, to scientific theories, medical issues, and political situations.

As opposed to highly curated sources like Wikipedia, unstructured social media contain the emotions, rumors, and more tentative connections between concepts. Beyond the factual repository, they record what is *interesting* to its users. While retrieval of information relevant to the user's need is always a core mission in web search, the current era of social media has shown the importance of moving beyond the factual information of the web to what society – and in particular one's social circle – considers as important and interesting, or, sometimes, just entertaining. The hours that users spend on Facebook, Twitter, Pinterest, and Instagram attest to the importance of socially-curated information to support an interesting and serendipitous search experience. Integrating this information into a search engine provides exciting new possibilities not only for the classic web search, but especially in the exploratory search – when the information need is loosely defined, and serendipity is welcome.

Our work [Bordino et al. 2013] demonstrates that unstructured social media represent a resource of paramount importance for supporting serendipitous discoveries in exploratory search. We achieve this by examining and comparing the potential of two extremely different sources of user-generated content – Yahoo Answers and Wikipedia, in promoting serendipitous search.

[1] `https://stackoverflow.com/`
[2] `https://www.healthtap.com/`

2. DEESSE (ENTITY-DRIVEN EXPLORATORY AND SERENDIPITOUS SEARCH SYSTEM)

We develop an entity-based exploratory search framework, dubbed DEESSE (entity-Driven Exploratory and serendipitous Search SystEm), and designed to support a serendipitous exploration of two popular social media: Wikipedia and Yahoo Answers. Yahoo Answers is nowadays the largest community question/answering portal, with millions of users posting hundreds of millions of questions and answers. Wikipedia is a popular collaboratively-edited and highly curated online encyclopedia. Although the un-curated nature of Yahoo Answers may make it a less trustworthy source of information, the freedom of conversation provides invaluable opportunities for interesting and serendipitous exploration.

DEESSE represents the content of each data source (Yahoo Answers and Wikipedia) as an entity network, and retrieves entities to a query entity using an algorithm based on lazy random walk with restart. This choice follows the latest research trend of adopting *Entity search* [Balog et al. 2010; Cheng et al. 2007] as the paradigm for extracting meaningful information items from huge volumes of data. Entity search bridges the two worlds of information retrieval and semantic web, combining the ability of the former to deal with the sheer size of web search data in a scalable manner, with the capacity of the latter, to discover and interpret complex relations among entities. Adopting a data model centered around entities, instead of documents, entity search (with prominent example Google's *Knowledge graph*) supports a direct, holistic search for targeted information concepts.

Following [Hauff and Houben 2012], we delve into what makes search serendipitous by using metadata accounting for the intensity of the emotion, the quality of the writing, and the topical category of the text surrounding each entity. We examine the extent to which each dimension contributes to the perceived serendipity.

We find that both Wikipedia and Yahoo Answers offer relevant results which are dissimilar to those found through a web search, and complementary in nature. However, Yahoo Answers shows to be better at favoring the most interesting entities, both when considering the query and personally to the labelers. The effects of constraining retrieval based on sentiment, quality or topic vary across datasets and topical categories, suggesting that it is not enough, for instance, to select only emotionally-evocative items in order to catch the user's interest. A demo of DEESSE is available online.[3]

3. THE SYSTEM

We briefly describe DEESSE. Full details can be found at [Bordino et al. 2013].

Data and Languages. DEESSE exploits two main data sources, Yahoo Answers and Wikipedia, which are both available in different languages. The current version of the system supports English and Spanish, building a separate network instance for each language and for each source. Our Yahoo Answers dataset consists of a dump of English and Spanish questions from 2010-2012, and the answers to these questions. For Wikipedia, we extracted the English and the Spanish dumps from December 2011.

[3] [http://deesse.limosine-project.eu.

Entity extraction. Fro each source we build a network of all the entities that occur in its documents. For entity extraction, we follow the common approach that for an extracted entity to exist, it must appear as a Wikipedia page. Our entity-extraction methodology was chosen due to its suitability for large-scale data processing. It uses a machine-learning approach proposed in [Zhou et al. 2010] for resolving surface forms extracted from the text to the correct Wikipedia entity, and Paranjpe's *aboutness* ranking model [Paranjpe 2009] to rank the obtained entities according to their relevance for the text.

Network Extraction. Given the set of entities extracted from a language-specific dataset, we construct a network using a content-based similarity measure to create arcs between entities. Adopting the vector-space model, we represent each entity by a TF/IDF vector, extracted by the whole set of documents where the entity appears. We then measure the similarity between any two entities by the cosine similarity of the corresponding TF/IDF vectors. The all-pairs similarity computation required to build the network is performed efficiently by using a distributed algorithm [Baraglia et al. 2010] that works on Hadoop.[4]

Entity feature extraction. Each entity network in the system is enriched with metadata regarding topic, quality and sentiment. Topical features are built exploiting a proprietary taxonomy to assign topical categories to the documents. We derive category features for each entity in a graph, by aggregating over all the documents where the entity appears, and retaining the top three categories that are most frequently associated with such documents.

To build sentiment features, we classify the originating documents with SentiStrength,[5] obtaining a positive and a negative score that we combine into a *sentimentality* score [Kucuktunc et al. 2012], measuring the global amount of sentiment. Entity-level sentimentality scores are obtained by performing the computation on small windows (20 words) of text around each mention of an entity, and then averaging across all mentions.

Quality is measured in terms of *readability*, the difficulty that a reader may encounter in comprehending a text. Lower readability scores are assigned to more sophisticated documents, which require higher education level to be understood. For each document we compute the Flesch Reading Ease score [Flesch 1948]. The readability score of any entity is the median Reading Ease over all the documents where the entity appears.

Entity Ranking. The ranking module extracts from a network, the top n entities that are most related to a query entity. Our method is inspired by random-walk algorithms [Jeh and Widom 2003; Tong and Faloutsos 2006], which are successfully applied in many recommendation problems. The algorithm performs a *lazy* random walk with restart to the input entity, ranks all nodes based on the stationary distribution of the walk, and selects the n entities with highest rank. We use an efficient `giraph`[6] implementation of random walks.

Constrained Retrieval. We use the metadata features computed for the entities to constrain the retrieval in the dimensions of sentimentality, quality, and topicality. This is attained by filtering the results of the original retrieval to retain the top entities that satisfy a given constraint (for instance, a given topic, or *high* or *low* sentimentality or quality).

[4] `hadoop.apache.org`
[5] `http://wwww.sentistrength.wlv.ac.uk`
[6] `http://giraph.apache.org`

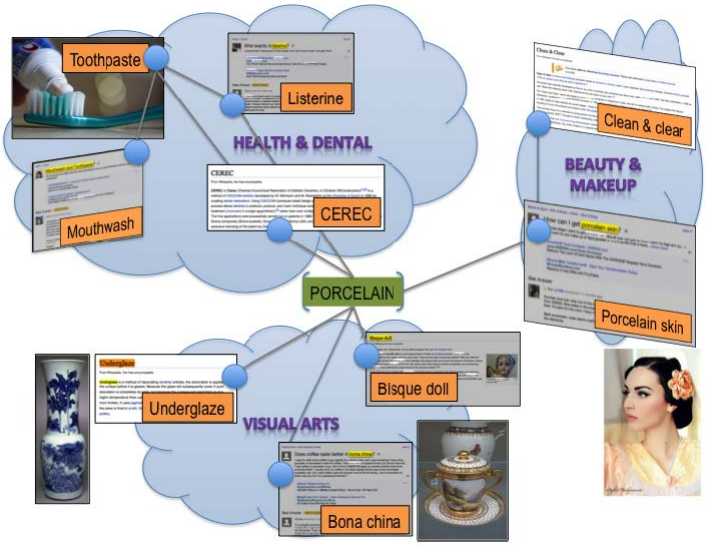

Fig. 1. Example of entities and bundles retrieved for the query entity "porcelain".

Bundled Retrieval. The current version of the system extends the original retrieval module by adopting the paradigm of *composite retrieval* [Mendez-Diaz et al. 2014], which recently emerged as a powerful way to assist the users with complex information seeking activities. The idea is to not show results in a ranked list, but instead organize them into item *bundles* designed to satisfy a number of properties, based on the users's preferences or needs. DEESSE implements a topical-bundling algorithm that organizes search results into topically coherent bundles, based on the categories of the query entity.

We finish with an example of an entity query "porcelain", for which DEESSE returns the entities and the bundles shown in Figure 1. The results are grouped into three bundles. "Health & Dental" covers entities such as *CEREC*, a method for dental restorations, and related entities for dental health like *Listerine* and *Toothpaste*. Next, in "Beauty & Makeup" the idea of obtaining what is called *porcelain skin* is captured, while "Visual Arts" contains art products that are actually made of porcelain, such as *Bisque doll* or *Underglaze*.

4. EVALUATION

Serendipity, or the act of unexpectedly encountering something unexpected and interesting, is widely regarded as "valuable in the processes of science, technology, art, and *daily life*" [Andre et al. 2009]. Discoveries have been made through serendipity, users happily divert from their original information needs when experiencing serendipity, and even recommender systems acknowledge the importance of serendipity in the type of user experience they wish to promote. Our work is concerned with serendipity as experienced in daily life: a user submitted a query to fulfill an information gap of hers, or to simply browse. Evaluating which search results trigger serendipity is however not an easy task.

Our work compares the search results extracted from the Yahoo Answers and Wikipedia

entity networks, or their versions constrained on sentiment, topic or quality, in terms of serendipity. We consider two approaches to assessing serendipity. The former, proposed in [Ge et al. 2010] in the context of recommender systems, considers two main attributes of serendipity: unexpectedness (or surprise) and usefulness (relevance). Unexpectedness can be computed by comparing to some "obvious" baseline. We take entities appearing in the top results of major commercial search engines as those which users may expect to see during the usual search experience. Usefulness can be estimated using standard information retrieval relevance judgments. The extent of serendipity in a set of search results can then be measured as the amount of results which are unexpected (they do not appear in the baseline runs), but also useful (relevant).

We take several peculiarities of our task into account, and consider several versions of this measure. First, because Wikipedia is often at the top of results returned by the search engines, we may want to exclude it from consideration, especially when evaluating Wikipedia-based tools. Second, we exploit the related query suggestions of some search engines, which use query logs and other information to disambiguate and expand the search space.

Our second approach for serendipity goes beyond relevance by also considering the *interestingness* of the results. Previously, Andre et al. [Andre et al. 2009] evaluated web search results in terms of their relevance and "interestingness", hypothesizing that "search results that are interesting but not highly relevant indicate a potential for serendipity." A more personal and subjective notion, it calls for an extensive manual evaluation. Because it is difficult to quantify this measure, we take an approach introduced by Arguello et al. [Arguello et al. 2011], which uses pairwise comparisons between all of the result pairs to build a reference result ranking. However, it is prohibitively expensive to label each possible pair, so to estimate the proper rank of a result we sample comparison pairs for each result from all possible permutations, and we use a voting methodology to rank them. We can then use a rank-based distance metric such as Kendalls tau-b, which takes into account the ties, as well as the ranking differences – the "closer" the result ranking is to the reference ranking, the better the tool estimates interestingness of its results.

We use the above measures to compare results extracted from Yahoo Answers and Wikipedia for 50 query entities, including people, places, events, websites, gadgets, sports, and health-related topics. Both measures – one based on surprise and the other on interestingness – show the Wikipedia-based system to be outperformed by that using Yahoo Answers [Bordino et al. 2013]. The first measure (fraction of unexpected recommendations which are also relevant) is typically 6% - 7% higher, with the difference even higher (10% - 15%) when results are filtered based on readability. For the second measure – similarity to a reference interestingness ranking– we use Crowdflower[7] to gather over 7,000 comparisons of results. When comparing to the ranking reflecting users' personal interest in a topic, we find the tau-b of the Yahoo Answers system to be at 0.324, compared to 0.139 of Wikipedia.

Beyond that, this procedure helps in discovering results which would be downgraded by traditional search engines, but which our method promotes – those with more interestingness, or with more surprise. Entities which may be not the most relevant, but quite interesting to the users may be books (such as the New York Times bestseller "Water for Elephants" for query "Robert Pattinson"), or educational articles (like on the History of

[7] http://www.crowdflower.com/

Ptolemaic Egypt for query "Egypt").

During this evaluation, we also test the constraining of the results based on metadata extracted from the text (sentiment, quality, and topics). We find the results mixed, and potentially highly dependent on the topic matter. For example, surprisingness improves when we favor the entities appearing in highly readable text, but interestingness decreases. We discuss the opportunities of using metadata for enriching exploratory search systems next.

5. THE POWER OF METADATA

The richness of user-generated content provides connections between entities which are built by people, reflecting both semantic and more non-obvious connections between them. The same content also provides the social, topical, spacial, and temporal context for each entity, recording metadata which may be of use during serendipitous search. Unlike structured and semi-structured descriptions of the various attributes of an entity, such as what can be found in knowledge bases or Wikipedia, metadata derived from social media provides a glimpse into the "life" of the concept in people's daily interactions.

Combining search results with their metadata, faceted search allows a user to explore the results from different angles, and provides a more engaging alternative for exploration to traditional query reformulation [Kules et al. 2009]. Research in faceted search has shown that users prefer to interact with multiple facets (even hierarchical ones), and that they feel in control of the experience [English et al. 2002]. This is why we consider several alternative metadata – topic, sentiment, and quality, which we mine from the content and associate with each entity, as a means of supporting serendipitous exploratory search.

In our work [Bordino et al. 2013], we investigate whether filtering search results based on the above metadata, improves serendipity and interestingness as perceived by the users. Our results are inconclusive. Indeed, we find that the effects of metadata constraints vary across datasets and topical categories, with no clear understanding as to the why. In any case, it is simply not enough, for instance, to select only emotionally-evocative items to catch the user's interest. In the following we describe the potential and the challenges of the metadata analyzed, and we also discuss some additional dimensions that we consider worth investigating to promote serendipity in exploratory search.

Topics. Topical hierarchies have been extensively used to structure search results. Data-driven dynamic approaches have been proposed to adjust the category facets to better suit the user's query [Tunkelang 2006]. Extracting the facet terms from the text itself has been proposed in [Dakka et al. 2006], where external resources (Wikipedia, Google, and WordNet) are used to generalize the terms used in the text. Since these facets may overlap, [Carterette and Chandar 2009] propose a set-based probabilistic model for reducing redundancy. As explained in Section 3, our system assigns three topical categories to each entity, and organizes the results retrieved from the network into topically coherent "bundles", based on the categories of the query entity. A crowdsourcing evaluation of this bundled retrieval approach showed that 77% of users preferred such topical bundles to a standard ranked list of results, for their usefulness in finding new topics to explore.

Sentiment. A popular cross-topic dimension, sentiment has been used to explore blogs [Fujimura et al. 2006], YouTube videos [Grassi et al. 2011], and Tweets [Walther and

Kaisser 2013]. Sentiment lexicons, such as SentiWordNet[8] and SentiStrength[9], have been developed using both manual and automatic efforts, and are commonly used for enriching social media. Our system (see Section 3) uses SentiStrength to associate a sentiment with each entity, applying the lexicon on small windows of text around each mention of an entity, and combining the resulting scores into a "sentimentality" score, measuring the extent to which an entity appears in highly emotional or opinionated contexts. We investigated [Bordino et al. 2013] whether returning entities conveying higher or lower emotion improves the serendipity and interestingness of results. The appropriateness of such information varies across topics: better results are obtained for those associated with emotional speech, such as sports. It also depends on the data source, with richer emotional content available on non-curated sites like Yahoo Answers benefitting more than if Wikipedia is used. Finally, sentiment analysis is far from perfect, and it is important to evaluate the performance of sentiment analysis tools to properly depict the aggregate view of the data.

Readability. A proxy for text quality, a measure of readability, or how well the text was written, has been used to examine diversity of the Web content [Kim et al. 2012]. Our system uses the Flesch Reading Ease metric [Flesch 1948], assigning higher scores to text that is easier to read. In our evaluation, filtering search results based on readability did not help in boosting the perceived accuracy or interestingness. We believe that the informal language of social media makes it difficult to use this metric to assess the quality of user-generated content. An alternative source could be user-generated metadata such as thumbs up/down, giving of points, or the reputation of the content's author.

Time. Adding a temporal dimension to categorical metadata, a timeline shows the major events and developments of the entity's history. Utilizing the graphical dimensions, time can be viewed as points, lines, cycles, branches, etc. [Aigner et al. 2007]. Visualizing information in two or three dimensions, users can interact with such a visualization by zooming in and out, rescaling the axes, rotating the plots, and getting more information in pop-ups [Tekusova and Schreck 2008]. The evolution of a topic across time has been an active research area for decades, but its application to exploratory search is yet to be tested.

Spacial data. Geographic mapping of data has become popular in Mashups [Wilde 2006] – a visualization that integrates various data sources and web-based applications to create a new application tailored to a specific task. A popular geographical representation application is Google Maps, described by [Miller 2006] as accessible, agile, adaptable, and data rich. The "virtual globe" of Google Earth has been used during the Hurricane Katrina to distribute information about the damage occurring in New Orleans [Crutcher and Zook 2009]. Yahoo Answers lacks a geographical dimension, but micro-blogging platforms such as Twitter, Facebook, Foursquare present a wealth of spacial information.

Language. As platforms open to a greater variety of users, content becomes available in a greater multitude of languages. For example, there are Wikipedia versions for 126 languages, and Yahoo Answers operates in 12 languages. Our system currently supports English and Spanish, building a separate network instance for each language and for each source. Support for cross-linguality is currently under development. Synchronizing data

[8] http://sentiwordnet.isti.cnr.it/
[9] http://sentistrength.wlv.ac.uk/

sources in different languages is a great challenge, but could unite not just language, but culture-specific values and knowledge surrounding the mentioned entities.

Demographics. Demographics such as gender and age, may be useful in assessing the population speaking on a topic. A search engine for emotional content, We Feel Fine [Kamvar and Harris 2011] provides such demographic breakdowns of sentences that include words like "I feel" or "I am feeling". When visualized using color, shapes, and volume, this data provides a glimpse into the change of emotions across these dimensions, showing, for example, that with age people express considerably less anger and sadness.

6. OPEN CHALLENGES AND CONCLUSION

We have presented DEESSE, an entity-based search system designed to support a serendipitous exploration of two popular social media: Yahoo Answers and Wikipedia. The system is based on networks of entities extracted from the documents, enriched with further metadata, still derived from the content, and regarding the intensity of the emotion, the quality of the writing, and the topical category of the text surrounding each entity.

Designing, developing, and evaluating a system for serendipitous exploratory search poses a number of critical challenges. Our work has faced and solved some of these, but many questions remain open. We discuss some of the most important ones below.

Improving entity relations with semantic information and social signals. A first critical improvement for our current approach would consist in extracting more informed entity relations. When building the entity network we now employ only syntactic measures by means of cosine similarity between the corresponding TF/IDF vectors representing documents. Similarity between two documents can be extended beyond this solely syntactic comparison by for instance adding semantics, or by combining different language domains. With respect to ranking entities, we envision the fusion of existing signals with social signals to obtain more relevant and diverse related entities for a given query entity. It would be also important to refine the definition of goodness (be it relevance or serendipity) by incorporating temporal characteristics of recency, trendiness, and novelty.

Large scale. Given the size of the datasets involved, current algorithms are pushed to their limits. Entity extraction and network building at scale requires distributed parallel computing infrastructure such as the Hadoop framework[10]. Our algorithm for retrieving entity search results from the networks consists in a random-walk computation, which even using an optimized parallel `giraph`[11] implementation, requires a computational cost of minutes per query entity. This cost makes running the ranking module at query time prohibitive. To make our solution viable, we perform the computation offline, and we store and index the resulting distributions, which are then retrieved at query time. We use state-of-the-art pruning and bucketing techniques to perform such operations efficiently. However, we see plenty of room for improving our current algorithms. An effort to reduce the computational costs could be extremely beneficial, as it would allow to update the system more frequently, making it able to serve more timely results.

[10] `hadoop.apache.org`
[11] `http://giraph.apache.org`

Multi-lingual, unstructured data. A testing ground is being created for the improvement of existing and development of new NLP tools. Indeed, performing the task of entity recognition on unstructured and heterogeneous social media arises additional difficulties with respect to those that must be solved when dealing with more structured sources or even standard web pages, due, for example, to the significantly smaller context and to the inherently multilingual nature of the data, which makes it no longer sufficient to only accept English as the dominant language. Tools like SentiStrength[12] are capable to extract sentiment in various languages, but many existing toolkits focus solely on English. Even if we arrive at the point that entity information and relations can be multilingually extracted, a whole new research field emerges on how to combine and present information about a given entity over different language domains.

Evaluation. When devising algorithms to drive the curiosity of users by promoting exploratory and serendipitous search, one enters a field of research that lacks objective metrics to measure the success of an approach. Formalizing user's experiences is mainly obtained by meticulously designing a user study. There is an unexplored area for research on establishing meaningful measures that are capable of capturing user curiosity, engagement, satisfaction with respect to both information need and serendipitous discoveries.

Our goal in building DEESSE was that of comparing Yahoo Answers and Wikipedia with respect to their ability in promoting a serendipitous and interesting exploratory search experience. Our extensive crowdsourcing evaluation showed that users prefer Yahoo Answers to Wikipedia for the possibility of discovering more surprising and interesting entities. Unstructured social media that allow freedom of conversation are increasingly becoming a powerful mirror of the collective view on various various economic, politic and cultural phenomena. Our results demonstrate that these media represent an extremely rich source to exploit for engaging users in entertaining and fruitful explorations. We believe that the cauldron of user-generated content provided by media such as Yahoo Answers could become a formidable ingredient to successfully promote serendipity in search. We hope that our work provides good pointers for further research in the field.

ACKNOWLEDGMENTS

This work was partially funded by the Linguistically Motivated Semantic Aggregation Engines (LiMoSINe[13]) EU project.

[12] http://wwww.sentistrength.wlv.ac.uk
[13] www.limosine-project.eu

REFERENCES

AIGNER, W., MIKSCH, S., MÜLLER, W., SCHUMANN, H., AND TOMINSKI, C. 2007. Visualizing time-oriented dataa systematic view. *Computers & Graphics 31,* 3, 401–409.

ANDRE, P., TEEVAN, J., AND DUMAIS, S. T. 2009. From x-rays to silly putty via uranus: Serendipity and its role in web search. *SIGCHI*.

ARGUELLO, J., DIAZ, F., CALLAN, J., AND CARTERETTE, B. 2011. A methodology for evaluating aggregated search results. In *ECIR*.

BAEZA-YATES, R. 2010. Searching the web of objects. In *Proceedings of the Third International Conference on Objects and Databases*. ICOODB'10. Springer-Verlag, Berlin, Heidelberg, 6–7.

BALOG, K., MEIJ, E., AND DE RIJKE, M. 2010. Entity search: Building bridges between two worlds. In *Proceedings of the 3rd International Semantic Search Workshop*. SEMSEARCH '10. ACM, New York, NY, USA, 9:1–9:5.

BARAGLIA, R., DE FRANCISCI MORALES, G., AND LUCCHESE, C. 2010. Document similarity self-join with mapreduce. In *ICDM*.

BORDINO, I., MEJOVA, Y., AND LALMAS, M. 2013. Penguins in sweaters, or serendipitous entity search on user-generated content. In *Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management*. CIKM '13. ACM, New York, NY, USA, 109–118.

CARTERETTE, B. AND CHANDAR, P. 2009. Probabilistic models of ranking novel documents for faceted topic retrieval. In *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 1287–1296.

CHENG, T., YAN, X., AND CHANG, K. C.-C. 2007. Supporting entity search: A large-scale prototype search engine. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*. SIGMOD '07. ACM, New York, NY, USA, 1144–1146.

CRUTCHER, M. AND ZOOK, M. 2009. Placemarks and waterlines: Racialized cyberscapes in post-katrina google earth. *Geoforum 40,* 4, 523–534.

DAKKA, W., DAYAL, R., AND IPEIROTIS, P. 2006. Automatic discovery of useful facet terms. In *SIGIR Faceted Search Workshop*. 18–22.

ENGLISH, J., HEARST, M., SINHA, R., SWEARINGEN, K., AND LEE, K. 2002. Flexible search and navigation using faceted metadata. Tech. rep., Technical report, University of Berkeley, School of Information Management and Systems, 2003. Submitted for publication.

FLESCH, R. 1948. A new readability yardstick. *Journal of Applied Psychology 32,* 3 (June), p221 – 233.

FUJIMURA, K., TODA, H., INOUE, T., HIROSHIMA, N., KATAOKA, R., AND SUGIZAKI, M. 2006. Blograngera multi-faceted blog search engine. In *International World Wide Web Conference, Proc. 3rd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*.

GE, M., DELGADO-BATTENFELD, C., AND JANNACH, D. 2010. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM, New York, NY, USA, 257–260.

GRASSI, M., CAMBRIA, E., HUSSAIN, A., AND PIAZZA, F. 2011. Sentic web: A new paradigm for managing social media affective information. *Cognitive Computation 3,* 3, 480–489.

HAUFF, C. AND HOUBEN, G.-J. 2012. Serendipitous browsing: Stumbling through wikipedia. *Searching 4 Fun*.

JEH, G. AND WIDOM, J. 2003. Scaling personalized web search. In *WWW*.

KAMVAR, S. D. AND HARRIS, J. 2011. We feel fine and searching the emotional web. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 117–126.

KIM, J. Y., COLLINS-THOMPSON, K., BENNETT, P. N., AND DUMAIS, S. T. 2012. Characterizing web content, user interests, and search behavior by reading level and topic. In *WSDM*.

KUCUKTUNC, O., CAMBAZOGLU, B. B., WEBER, I., AND FERHATOSMANOGLU, H. 2012. A large-scale sentiment analysis for yahoo! answers. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*. WSDM '12. ACM, New York, NY, USA, 633–642.

KULES, B., CAPRA, R., BANTA, M., AND SIERRA, T. 2009. What do exploratory searchers look at in a faceted search interface? In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 313–322.

MENDEZ-DIAZ, I., ZABALA, P., BONCHI, F., CASTILLO, C., FEUERSTEIN, E., AND AMER-YAHIA, S. 2014. Composite retrieval of diverse and complementary bundles. *IEEE Transactions on Knowledge and Data Engineering 99,* PrePrints, 1.

MILLER, C. C. 2006. A beast in the field: The google maps mashup as gis/2. *Cartographica: The International Journal for Geographic Information and Geovisualization 41,* 3, 187–199.

PARANJPE, D. 2009. Learning document aboutness from implicit user feedback and document structure. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. CIKM '09. ACM, New York, NY, USA, 365–374.

TEKUSOVA, T. AND SCHRECK, T. 2008. Visualizing time-dependent data in multivariate hierarchic plots-design and evaluation of an economic application. In *Information Visualisation, 2008. IV'08. 12th International Conference*. IEEE, 143–150.

TONG, H. AND FALOUTSOS, C. 2006. Center-piece subgraphs: Problem definition and fast solutions. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '06. ACM, New York, NY, USA, 404–413.

TUNKELANG, D. 2006. Dynamic category sets: An approach for faceted search. In *ACM SIGIR*. Vol. 6.

WALTHER, M. AND KAISSER, M. 2013. Geo-spatial event detection in the twitter stream. In *Advances in Information Retrieval*. Springer, 356–367.

WILDE, E. 2006. Knowledge organization mashups. *Retrieved 8*, 2007.

ZHOU, Y., NIE, L., ROUHANI-KALLEH, O., VASILE, F., AND GAFFNEY, S. 2010. Resolving surface forms to wikipedia topics. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1335–1343.

Ilaria Bordino is a Research Scientist at Yahoo Labs, Barcelona. She works on Web Information Retrieval, Web Mining and query-log Mining. Her current interests include (but are not limited to) analyzing web and social-media data to track real-world events and trends (including applications to the financial domain); studying how to improve the web users' search and exploratory experience by recommending relevant and interesting information.

Mounia Lalmas is a Principal Research Scientist at Yahoo Labs, London. Her work focuses on developing models and metrics of user engagement, through the study of user behavior, web analytics, the analysis of users emotion and attention, and mouse and gaze movement. She is studying user engagement in areas such as advertising, digital media, social media, and search, and across devices (desktop, tablet and mobile). She also pursue research in social media and search.

Yelena Mejova is a scientist at the Qatar Computing Research Institute, a part of the Social Computing group. Specializing in text retrieval and mining, Yelena is interested in building tools for tracking real-life social phenomena in social media. Her work focuses on sentiment classification and large-scale user-generated text analysis, and in particular tracking social phenomena including political opinion, poll now-casting, public health, and web-mediated charitable giving and international communication.

Olivier Van Laere is a Research Engineer at Yahoo Labs, Barcelona. He works on Web scale data mining, in particular from social media. His research interests are currently focused around mining geographical clues from social media (e.g. automated geotagging), large scale online classification, mining financial data and exploratory search.