


## Beyond entities: promoting explorative search with bundles

Ilaria Bordino<sup>1</sup>  · Mounia Lalmas<sup>2</sup> · Yelena Mejova<sup>3</sup> · Olivier Van Laere<sup>4</sup>

Received: 7 March 2015 / Accepted: 8 June 2016 / Published online: 13 July 2016  
© Springer Science+Business Media New York 2016

**Abstract** Search engines are increasingly going beyond the pure relevance of search results to entertain users with information items that are interesting and even surprising, albeit sometimes not fully related to their search intent. In this paper, we study this serendipitous search space in the context of *entity search*, which has recently emerged as a powerful paradigm for building semantically rich answers. Specifically, our work proposes to enhance an explorative search system that represents a large sample of Yahoo Answers as an entity network, with a result structuring that goes beyond ranked lists, using *composite entity retrieval*, which requires a *bundling* of the results. We propose and compare six bundling methods, which exploit topical categories, entity specializations, and sentiment, and go beyond simple entity clustering. Two large-scale crowd-sourced studies show that users find a bundled organization—especially based on the topical categories of the query entity—to be better at revealing the most useful results, as well as at organizing the results, helping to discover novel and interesting information, and promoting exploration.

---

*Data availability* To ensure repeatability of our experiments, we make our entity networks and the results of our user studies available upon request.

---

✉ Ilaria Bordino  
ibordino@acm.org

Mounia Lalmas  
mounia@acm.org

Yelena Mejova  
ymejova@qf.org.qa

Olivier Van Laere  
olivier@getblueshift.com

<sup>1</sup> Unicredit, R&D, Rome, Italy

<sup>2</sup> Yahoo Labs, London, UK

<sup>3</sup> Qatar Computing Research Institute, Doha, Qatar

<sup>4</sup> Blueshift Labs, San Francisco, CA, USA

Finally, a third study of 30 simulated search tasks reveals the bundled search experience to be less frustrating and more rewarding, with more users willing to recommend it to others.

**Keywords** Entity search · Entity networks · Composite retrieval · Bundles · Explorative search · Topical bundles

## 1 Introduction

The classic Web search experience, consisting of 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 heterogeneous sources on and off the Web. Having to face a *web of objects* (Baeza-Yates 2010) rather than a web of links, 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.

The increasing need for supporting expressive and yet intuitive querying over large-scale, heterogeneous and unstructured data has led to a surge in research on *entity search* (Balog et al. 2010; Cheng et al. 2007; Hoffart et al. 2011; Kulkarni et al. 2009; Mihalcea and Csomai 2007; Milne and Witten 2008; Paranjpe 2009), along with evaluation efforts such as INEX Entity and Linked Data tracks,<sup>1</sup> TREC Entity track,<sup>2</sup> and SemSearch challenge.<sup>3</sup>

In the past years, entity search has emerged as a prominent alternative to document search, and an ideal paradigm to support *exploratory search* activities, because it provides semantically rich answers, such as entities (people, places, events) and their relations, which in these scenarios are often considered more suitable for search exploration than individual web pages. Unlike the familiar query-driven search paradigm in which a relevant set of documents are sought after, exploratory search addresses a problem of less well-defined information need. It considers a scenario that has elements of uncertainty, because the information seeker is unfamiliar with the problem domain, or the search task requires some exploration (White et al. 2008). Often, users enjoy exploring without a specific search objective in mind, but rather with a simple desire to get an update, or be entertained during their spare time. In this scenario, *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. In our work, we consider a particular kind of exploratory search, one which is driven by a particular need, but in which complexity (of the information need) and flexibility (exploration is welcomed) are present. Following (Miliaraki et al. 2015), we dub this task as *explorative search*, as one driven by suggestions of the system to the initial interest of the searcher.

<sup>1</sup> <http://www.inex.otago.ac.nz/tracks/entity-ranking/entity-ranking.asp>.

<sup>2</sup> <http://ilps.science.uva.nl/trec-entity/>.

<sup>3</sup> <http://semsearch.yahoo.com/>.

Finally, *serendipitous search* occurs when a user with no a priori or totally unrelated intentions interacts with a system and acquires useful information (Toms 2000). Serendipity—a discovery of something new and interesting—has been an important consideration for recommender systems (Iaquinta et al. 2008), and is becoming increasingly important for search systems, which are now often constructed, at least partly, with the objective of engaging users, so as to keep them interacting with the website even beyond a predefined purpose. This is done for a number of reasons, including encouraging learning and unexpected discoveries, supporting a certain kind of shopping behavior, and exposing users to advertisements.

*How, then, could we present heterogeneous, semantically rich results such as those returned by an entity search system so as to encourage interaction and further exploration?*

In this paper we investigate how to answer this question, using Yahoo Answers as a case study. While Yahoo Answers may not be optimal for factoid search (Liu and Agichtein 2008), it is a destination for complex information needs such as opinion or advice. As opposed to highly curated sources like Wikipedia, unstructured social media like question/answering fora contain the emotions, rumors, and more tentative connections between concepts. Beyond the factual repository, they record what is *interesting* to its users. Integrating this information into a search engine provides exciting new possibilities not only for the classic web search, but especially in the explorative search—when the information need is loosely defined, and serendipity is welcome.

We begin with a system (Bordino et al. 2013a, 2014; Laere et al. 2014) that we developed to support explorative entity search on Yahoo Answers.<sup>4</sup> The system represents Yahoo Answers data in the form of an entity network, with entities extracted from the documents (in this case, questions and answers) and connected according to the similarity of the texts in which they appear. The network is further enriched with metadata that provide additional dimensions for the analysis, such as topical categories and the sentiment expressed in the conversation. The system supports entity search by employing a random-walk based method to retrieve relevant entity results for a given entity query.

In our previous work (Bordino et al. 2013a, 2014) we showed that users value Yahoo Answers better than a more curated and trustworthy data source, such as Wikipedia, when it comes to the possibility of discovering interesting entity results. Our study showed that both media offer relevant results that are dissimilar to those found through a web search, but are complementary in nature. However, Yahoo Answers was preferred for favoring the most interesting and serendipitous results. We now go one step further, studying how to organize the results provided by our entity search tool to encourage interaction and further exploration.

For the above purpose, we extend our entity search system to support *composite retrieval*. Recent research (Amer-Yahia et al. 2014) proposes this novel paradigm as a way to assist the users with complex information seeking activities that involve running multiple search queries. For example, planning a trip may require gathering information about different places, reading online reviews to find hotels, and checking geographical proximity of places to visit.

Instead of providing the user with a traditional ranked list of information items, composite retrieval proposes to organize results into item *bundles* or groups designed to satisfy various properties, such as internal cohesiveness and external diversity, or other application-specific constraints. For example, when planning a trip, the user might have a limited

<sup>4</sup> A demo of the tool is available at <http://deesse.limosine-project.eu>.

financial budget or a fixed number of days available, and she might want to see different kinds of attractions during the visit to a new city (museums, parks, shops). In this scenario, a good composite answer should consist of bundles proposing alternative sets of activities that include different attractions, respect the budget constraint, and are compatible by geographical proximity.

In many ways, composite retrieval on the Web is similar to category-based search result clustering, which has been studied extensively in previous work (Ferragina and Gulli 2004; Käki 2005; Stefanowski and Weiss 2003; Wilson et al. 2010), showing that hierarchical presentation of results improves navigation of results and is more effective, in terms of search time, exploration of results and discovery of content, than traditional ranked lists. Our research aims at investigating what contribution composite retrieval can bring in the context of entity-driven *explorative* search. We are particularly interested in understanding whether explorative entity search can be extended with composite retrieval to achieve an improved user experience. To do so, we extend our original entity search tool with the capability of organizing the entity results retrieved for a query entity, into bundles.

We propose six bundling algorithms, which respectively group the result entities extracted as answer to a query based on (1) the topical categories of the query entity, (2) the categories shared by the result entities, (3) the sub-topics of the query entity (identified using search-log data), and (4) the aggregated sentiment of the documents in which the entity is present, as well as in conjunction with topical dimension (5, 6). Using different metadata and requiring constraints of different complexity, the algorithms vary in the way they explore the network and the initial result ranking of the tool. Our aim was not to exhaustively explore all the possible bundling approaches, but to study how returning bundled entities compares to returning entity lists in promoting explorative search. We nonetheless experiment with several bundling approaches to ensure that our results do not depend on the specifics of one approach.

We conduct an extensive experimental evaluation of the proposed bundling algorithms. Two large crowd-sourced studies are performed to first compare the bundles to a ranked-list baseline and next to identify the best bundling method. In a third study, we then use our interactive prototype to simulate an explorative search experience and gather more qualitative feedback on the performance of the bundling algorithms. In all three studies we employed Amazon Mechanical Turk, collecting thousands of annotations.

The first two studies, conducted on 150 popular search queries, showed that the bundling algorithm using the topical categories of the query entity is overwhelmingly preferred by the users over the standard ranked list. This method won all the comparisons with the other bundling algorithms and with the baseline ranked list, collecting up to 77 % of the preferences for some of the questions asked. We then compared the winning method to the typical ranked list in more complex information seeking activities. We devised 30 simulated explorative tasks, and asked users to express their preferences through questions pertaining to search effectiveness, user involvement, perceived usability, and task endurance. The result bundles based on the categories of the query were preferred to the standard ranked list for providing good overviews and facilitating the finding of relevant information; they were considered more involving and less frustrating, and worth recommending.

As a first foray into the application of composite retrieval to general-purpose explorative entity search, this work is a contribution to the overall enriched-search literature.

## 2 Related work

As the Web has made available a huge variety of textual and multimedia resources, people have increasingly started to perform complex search tasks, aimed at finding rich answers from various data sources. To satisfy these complex information needs, search systems are required to build composite solutions that aggregate information items according to various constraints and quality criteria. The concept of responding to information retrieval queries by presenting a composition of items has been investigated by many (Cao et al. 2011; Guo et al. 2012; Kashyap and Hristidis 2012; Parameswaran et al. 2011; Tran et al. 2011). In particular, our study is inspired by a recent study (Miliaraki et al. 2015) which evaluates the *explorative entity search* paradigm on a web search engine, expanding the setting beyond the standard web search.

At its simplest, composite retrieval is akin to category-based search-result clustering (Ferragina and Gulli 2004; Käki 2005; Stefanowski and Weiss 2003; Wilson et al. 2010). Categorizing search results is not new, specifically in the context of exploratory search (White and Roth 2009). For example, Yee et al. (2003) apply a categorizing approach to an image search and browsing task. Similarly, hierarchical categories have been used (Chen and Dumais 2000; Pratt and Fagan 2000; Wu et al. 2003) to show that the concept hierarchies provide easy navigation and outperform the typical ranked-list interface in search time and discovery. Scatter/Gather (Cutting et al. 1992) presents users with automatically computed summaries of clusters of similar documents and allows them to navigate through these summaries at different levels of granularity. Compared to the standard ranked result list (Pirolli et al. 1996), Scatter/Gather has been shown to induce a more coherent conceptual image of a text collection and communicate the distributions of relevant documents in the collection. Finally, a study of Käki (2005) reveals that categorized results can help users find useful or interesting items when document ranking fails. We complement these studies and investigate how entity search can promote explorative search within the paradigm of composite retrieval.

As opposed to web search where results are web pages, entity search provides a more semantically cohesive view of information with results being people, organizations, places, etc. The problem of discovering interesting relations from unstructured text has led to a surge in research on entity search (Hoffart et al. 2011; Milne and Witten 2008; Paranjpe 2009). To extract entities from raw text, the common approach (one which we adopt) is to map text to a Wikipedia page, which signifies an entity. In our work (Bordino et al. 2013a), given a search query, we retrieve other entities relevant to it by first building an entity network (Chakrabarti et al. 2006; Cheng et al. 2007) based on a pairwise entity relevance score, and by then applying random-walk computations on this network (Craswell and Szummer 2007; Jeh and Widom 2003).

Other works (Amitay et al. 2009; Yogev et al. 2012) have proposed to build more complex entity-relationship models to support various types of search over entities and their relations. However, we believe that a graph of pairwise relations, which express a more general notion of *relatedness* that can be quantified automatically and in a variety of topical domains, is a more natural choice to model entity similarity in our context. Also, the aforementioned semantically richer models require the usage of structured query languages, whereas we target non-expert, every-day users, and do not wish to rely on a particular visualization paradigm. The design of our graph-based system is described in detail in (Bordino et al. 2013a; Laere et al. 2014) and summarised in Sect. 3.

Applying topic-specific composite retrieval to entity search, Angel et al. (2009) study the problem of querying documents to build packages formed by multiple entities. For example, a trip consists of a composition of entities such as a city, and hotel and flight recommendations. Later on, Roy et al. (2010) explore the idea of retrieving *bundles* of items in the form of a *star*, for example, an iPhone and all its accessories. Further, Deng et al. (2012) study the problem of recommending item bundles that satisfy multiple selection criteria and compatibility constraints. They introduce functions to compute the cost and usefulness of items to a user, and propose query generalizations to help users revise their criteria, when no sensible suggestion can be found. Recently, Bota et al. (2015) found that when creating bundles manually, users prefer relevant, diverse, and cohesive bundles, often centering them around a pivot document signifying a particular subtopic. In this work we operationalize these constraints using topic and sentiment metadata, following from our work that shows that these metadata bring different angles, some more successfully than others, in promoting explorative search (Bordino et al. 2013a, 2014).

User search behavior and motivation have been investigated at length (Jansen and Pooch 2001; Rose and Levinson 2004; Spink et al. 2002), but the design of explorative search and more generally exploratory search systems continues to develop. For instance, Wilson et al. (2010) provide a guide for designing future web search systems that preserve the taxonomy of results, whereas Yue et al. (2012) perform a user study in collaborative exploratory web search, outlining the main activities that such a system needs to support. Our work contributes to these efforts, providing a basis for a clustered data visualization and a framework for its evaluation in the context of entity search.

Finally, a work closely related to ours is that of Bota et al. (2014), who study composite retrieval in the context of aggregated search—where results from different verticals available on the Web (image, video, news) are returned to users. They develop several algorithms, treating relevance as their main criteria to construct bundles, and cohesion and diversity as secondary. To tackle the challenges arising from the heterogeneous nature of the data, they exploit entities to link relevant results across verticals. They also incorporate query intent into the formation of bundles. Our work complements theirs, as we focus on entity search and investigate how composite retrieval promotes exploratory search, more precisely entity-driven explorative search, in this context.

### 3 Entity network

Our initial explorative search system, built in previous work (Bordino et al. 2013a; Laere et al. 2014), consists of an entity network extracted from Yahoo Answers. A study reported by Liu and Agichtein (2008) suggests that while Yahoo Answers is not optimal for factoid search, it is becoming a popular destination for complex information needs such as opinion or advice. Moreover, we have shown (Bordino et al. 2013a) that users value Yahoo Answers more than Wikipedia for the possibility of discovering interesting results. Hence, this paper investigates further this potential.

#### 3.1 Dataset

The question and answering web portal, Yahoo Answers, allows people to ask questions on different topics and answer questions asked by other users, sharing their knowledge and opinions. Every question is assigned by the asker to one category in a hierarchy of

categories. Our initial dataset consists of a sample of Yahoo Answers documents from 2010 to 2011, containing English-language questions, and the answers to these questions.

### 3.2 Entity extraction

To extract Wikipedia entities from each document (question or answer), we parse each text to identify candidate mentions of Wikipedia entities. We then mark each mention with entity candidates retrieved from an offline Wikipedia database, and subsequently choose the correct entity by applying a resolution model (Zhou et al. 2010). We then use Paranjpe's *aboutness* ranking model (Paranjpe 2009) to rank entities according to their aboutness for the text, i.e. their goodness in representing a succinct representation of the main topic matters of the document.

Determining the aboutness (Paranjpe 2009) or salience (Gamon et al. 2013) of entities in web pages has become crucial for commercial search engines. In many cases only a small subset of entities is important for a given page, and considering entities that are poorly relevant to the main topic matter of a page can lead to degraded search experiences in an entity-triggered scenario. It should be emphasized that the concept of amounts or saliency differs from both the notions of importance and relevance. Importance refers to the general relevance of an entity outside the scope of the document. For example, an entity can represent a very famous and worldwide known personality, who can be peripheral to the specific subject matter of a document. At the same time, the relevance of an entity to the reader information need is something inherently subjective.

Our explorative system is built on a collection of Yahoo Answers questions and answers, extracting entities from each document and connecting entities based on the textual similarity of the text fragments they appear in. We employ Paranjpe's aboutness model on each input document (question or answer) to rank the entities occurring in it according to their saliency for the text, and discard those that are marginal to the page. Such poorly salient entities would just induce spurious low-weight arcs in the network. Paranjpe's model exploits structural and visual properties of web documents, and user feedback derived from search-engine click logs. The method achieved 75 % accuracy when evaluated against a ground truth of editorial relevance judgements for a collection of query-url pairs.

Although in the work of Paranjpe, the focus is on the detection of key term in web pages and not on entities, the keyword extraction task can be seen as related to salient entity extraction, where keywords and key phrases are a superset of salient entities in a document. This technique was state of the art when we built the main infrastructure of our explorative entity search tool (Bordino et al. 2013a). More advanced techniques have been successively developed to identify salient entities instead of salient terms, like the work of Gamon et al. (2013). The choice of the tools used to extract an entity network from our Yahoo Answers dataset was guided by the necessity of processing large-scale data, and their high effectiveness in our context. It is possible that replacing each module of the pipeline with more recent and advanced techniques (e.g. TagMe,<sup>5</sup> Babelfy,<sup>6</sup> Dexter<sup>7</sup>) or changing Paranjpe's model with a more refined aboutness ranking model tailored to entities, could lead to improved overall performance, this is not the aim of the present paper; thus we leave it for future work.

---

<sup>5</sup> <http://tagme.di.unipi.it/>.

<sup>6</sup> [www.babelfy.org](http://www.babelfy.org).

<sup>7</sup> <http://dexter.isti.cnr.it/>.

**Table 1** Basic characterization of the network extracted from Yahoo Answers

| # Nodes | # Arcs      | # Isolated | Avg degree | Max degree | # Largest CC      |
|---------|-------------|------------|------------|------------|-------------------|
| 896,799 | 112,595,138 | 69,856     | 251.10     | 231,921    | 826,402 (92.15 %) |

**Entity Similarity.** Using the above methodology (and tools), we extract 896,799 distinct entities. With these, we construct an entity network using a content-based similarity measure to create arcs between entities. Adopting the vector-space model (Salton et al. 1975), we represent each entity by a TF/IDF vector, extracted by the concatenation of all the documents where the entity appears. We measure the similarity between any two entities in terms of the cosine similarity of the corresponding TF/IDF vectors. Because the TF/IDF weights cannot be negative, the similarity values will range from 0 to 1. We create an undirected network by computing all the pairwise similarities between the entities, using an efficient distributed algorithm (Baraglia et al. 2010) that works on Hadoop.<sup>8</sup> To avoid considering poorly significant relations, we prune all the arcs with similarity lower than a minimum threshold  $\sigma = 0.5$ .<sup>9</sup>

The result is a network containing 896,799 nodes and 112,595,138 arcs. Table 1 reports some basic characterization statistics about the network, listing number of nodes, number of arcs, number of isolated nodes, average and maximum degree, and the size of the largest connected component. The graph has a giant connected component spanning 92.15 % of the nodes. This is due to the presence of ultra-popular entities, representing very common concepts that appear ubiquitously in the dataset.

## 4 Bundling methods

The bundling algorithms presented in this paper extend the original retrieval algorithm of our system (Bordino et al. 2013a), which, given a query entity, returns a ranked list of result entities. We describe the original algorithm, as it is a component of our bundling methods, and also serves as a baseline to compare them. Next we present the bundling algorithms.

### 4.1 Baseline non-bundling algorithm

Our original retrieval method is inspired by random-walk based algorithms (Jeh and Widom 2003; Tong and Faloutsos 2006), which have been successfully applied in many recommendation problems (Bonchi et al. 2012; Bordino et al. 2013b; Craswell and Szummer 2007). The algorithm, implemented in `giraph`,<sup>10</sup> performs a random walk with restart to the input entity, following the links with probability proportional to the arc weights, and ranking all nodes based on the stationary distribution of this walk.

The algorithm applies two corrections to reduce unwanted bias towards popular entities with very large degree, which appear ubiquitously in the prominent positions of the ranking

<sup>8</sup> [www.hadoop.apache.org](http://www.hadoop.apache.org).

<sup>9</sup> The value of the threshold was chosen heuristically.

<sup>10</sup> <http://giraph.apache.org>.



vectors of all entities. First, we measure the rarity of any entity by computing its inverse document frequency. Given the ranking vector of an input entity, we filter out the top 500 entities with the lowest inverse document frequency (the value of this threshold was chosen heuristically). Second, we divide the ranking vector by the squared root of the global PageRank values obtained with no personalization, that is, (re-)starting at any node with uniformly random probability.

Our baseline retrieval algorithm is intended to allow users to explore the entity network by providing them with the entities that are most similar to the entity they are currently focusing on (the input entity). Behavioral data could be exploited to derive useful features for personalizing results based on user profile and activities. For example, in our framework based on Yahoo Answers, we could look at the browsing and contributing history of the users, the topics declared in their profile, comments, starts, thumbs up and down and so on. We plan to tackle this issue in the future.

In previous work, we editorially assessed the performance of this algorithm on 50 queries, reporting an average precision of 72.4 % with respect to *relevance* of results. This accuracy value is comparable with those achieved in other recommendation problems (Bonchi et al. 2012; Bordino et al. 2013b). We now extend these 50 queries with another 100, leading to a test set of 150 queries used to evaluate our bundling methods. All queries were sampled among the most searched queries in 2010/2011 from Google Zeitgeist. Although we could have used the logs of Yahoo search and Yahoo Answers, we turned to Google Zeitgeist to identify a set of publicly available popular queries to use for testing. This was to facilitate large-scale experiments with everyday users and different social media (our initial work compared Yahoo Answers and Wikipedia), and to favor the reproducibility of our experiments. The Zeitgeist queries were manually mapped to entities (Wikipedia articles) and filtered by coverage in the dataset, retaining those mentioned in at least 50 questions/answers. We then randomly sampled from the remaining queries. The resulting 150 queries are listed in Table 2.

## 4.2 Bundling algorithms

Through the above process, we maintain metadata of the entity network nodes to enrich it with topic, quality and sentiment features. We previously found that the topic metadata contributed most to improve performance in terms of the interestingness and relevance of search results. Based on this, and on the fact that a topical organization of the results is a natural choice to facilitate the exploration of a large-scale knowledge base, we focus mainly on topical bundles. We also experiment with sentiment as a criterion to create bundles, both on its own and in conjunction with topicality.

### 4.2.1 Topic and sentiment metadata

**Entity categories or super-topics.** In Yahoo Answers, each question is assigned one category chosen by the asker. Every answer to a question is listed under the category of the question. This manual topical classification is meant to help answerers, who typically find questions by browsing or searching the category hierarchy. We associate each entity in the graph with the top 3 categories that are most frequently assigned to the documents where the entity appears. Categories in the Yahoo Answers category may have up to three levels. The top-level categories are reported in Table 3. We refer to them as super-topics, to differentiate them from sub-topics, as introduced next.

**Table 2** Our test set of 150 queries

|                           |                                |                                  |                              |
|---------------------------|--------------------------------|----------------------------------|------------------------------|
| 1G                        | 2010 Haiti earthquake          | Adapa                            | Adele (singer)               |
| Amazon kindle             | American civil war             | Animal euthanasia                | Appendicitis                 |
| Aramaic language          | Asian studies                  | Asperger syndrome                | Talking point                |
| Bailout                   | Begging                        | Black Butler                     | Braxton Hicks contractions   |
| Bribery                   | Stefy                          | Carpentry                        | Cassandra                    |
| Chamomile                 | Chewing gum                    | Chickpea                         | Childbirth                   |
| Chile                     | Chinese people                 | Cholera                          | Cricket                      |
| Dallas Mavericks          | Daylight saving time           | Diary                            | Dieting                      |
| Dressage                  | Drooling                       | Earthquake                       | Egypt                        |
| Eiffel Tower              | Electro-magnetic radiation     | Eminem                           | Enrique Iglesias             |
| Essential fatty acid      | Stir frying                    | Ethanol                          | Euthanasia                   |
| Evaporation               | FIFA                           | FL Studio                        | Facebook                     |
| Fallopian tube            | Family medicine                | Fast food restaurant             | Game Boy Advance             |
| Genealogical DNA test     | Gluten                         | Graffiti                         | Greenhouse gas               |
| Haiti                     | Hard rock                      | Henna                            | Hernia                       |
| Honda accord              | Hound                          | Health hazard evaluation program | IPad                         |
| iPhone                    | Ice cube                       | Image stabilization              | Indigenous Australians       |
| Influenza                 | Influenza A virus subtype H1N1 | Jeggings                         | Jose Mourinho                |
| Justin Bieber             | Katy perry                     | Kim Kardashian                   | Kofi Kingston                |
| Lady Gaga                 | Larva                          | Leaf vegetable                   | Legal drinking age           |
| Libya                     | Linen                          | Llama                            | Loaf                         |
| Major depressive disorder | Matt Goss                      | McDonald's                       | Miami heat                   |
| Microfinance              | Microorganism                  | Middle East                      | Miley cyrus                  |
| Mobile phone              | Mount Everest                  | Natural gas                      | NASA                         |
| Netflix                   | New York Jets                  | Nicki Minaj                      | Nobel prize                  |
| Oil spill                 | Olympic games                  | Omega                            | Omnivore                     |
| Osama bin Laden           | Oxford street                  | Pain tolerance                   | Pap test                     |
| Parsley                   | Pedicure                       | Pertussis                        | Photosynthesis               |
| Plywood                   | Poland                         | Porcelain                        | Presidency of George W. Bush |
| Property tax              | Purgatory                      | RadioShack                       | Robert Pattinson             |
| Ron Paul                  | Sasuke Uchiha                  | Saul                             | Sean combs                   |
| Selena gomez              | Senior citizen                 | Shakira                          | Somnolence                   |
| Spanish empire            | Steve jobs                     | Subprime mortgage crisis         | Sulfuric acid                |
| Talk show                 | Tanning                        | Tennis                           | Terrorism                    |
| The jungle                | Thunderstorm                   | Touchpad                         | Trade union                  |
| Tsunami                   | Tux                            | Urology                          | Vaccine                      |
| Variety store             | Vedas                          | Venus                            | Vitamin D                    |
|                           |                                | Wayne Rooney                     | Wenger                       |

**Table 3** Yahoo! Answers top-level categories

|                         |                         |                         |
|-------------------------|-------------------------|-------------------------|
| Arts and humanities     | Beauty and style        | Business and finance    |
| Cars and transportation | Computers and Internet  | Consumer electronics    |
| Dining out              | Education andreference  | Entertainment and music |
| Environment             | Family andrelationships | Food and drink          |
| Games and recreation    | Health                  | Home and garden         |
| Local businesses        | News andevents          | Pets                    |
| Politics and Government | Pregnancy and parenting | Science and mathematics |
| Social science          | Society andculture      | Sports                  |
| Travel                  | Yahoo! Products         |                         |

**4.2.1.1 Entity specializations or sub-topics** The Yahoo Answers categories—as used above—are very general, and we can interpret them as the super-topics which an entity belongs to. The entities in our system represent concepts with different semantic granularity, where some cover multiple and diverse aspects. We therefore decided to further explore the topical dimension of our entities by identifying the sub-topics of an entity.

We apply an idea proposed by Capannini et al. (2011) to diversify web search results. Their method identifies the possible search intents behind a query, observing that whenever users are not satisfied with the results returned by a search engine for their query, they rephrase the query to provide a better formulation of their intent. The more specific reformulations of a query, which are called *specializations* (Boldi et al. 2009) and are generally obtained by adding words, are interpreted as possible *intents* or *sub-topics* of the query.

We adapt this idea to our entity search scenario, and exploit the wisdom of crowds provided by search-engine logs to identify *entity specializations*, i.e. entities representing more specific concepts, which we interpret as refinements or sub-topics of an entity. We process a large anonymized sample of the Yahoo search log, spanning the same time frame as our dataset. We apply the *query-flow* graph method (Boldi et al. 2008, 2009) to extract query-to-specialized-query transitions. We map the query-to-specialized-query transitions onto entity-to-specialized-entity transitions, extracting Wikipedia entities from each search query with a tool (Meij et al. 2012) optimized for the processing of very short texts. We weight each entity-to-specialized-entity transition by the aggregated frequency of all the originating query transitions.

**4.2.1.2 Sentiment (polarity)** A popular inter-topical 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<sup>11</sup> and SentiStrength<sup>12</sup> are commonly used for enriching social media. In our previous work—returning a ranked list of result entities that promoted serendipity (Bordino et al. 2013a)—with such a tool, we measured the extent to which an entity appears in highly emotional or opinionated contexts and used this information to rank the entities for a given query. We found that the appropriateness of such information depends on the topic, and in particular when associated with emotional speech, such as sports. We also found that it helped with richer emotional content available on non-curated sites such as Yahoo Answers (used in this work), compared to Wikipedia.

<sup>11</sup> <http://sentiwordnet.isti.cnr.it/>.

<sup>12</sup> <http://sentistrength.wlv.ac.uk/>.

In this work, therefore, we further explore sentiment to build bundles. To derive sentiment scores for entities, we classify the originating Yahoo Answers documents with SentiStrength, a state-of-the-art tool for short informal texts (such as the questions and answers from Yahoo Answers used here). The evaluation of the tool reported by the authors<sup>13</sup> shows correlation with human annotators at 0.55 and 0.56 for positive and negative sentiment scores, respectively. A topic-specific approach would certainly improve the sentiment classification, however in our entity-based system, this would require to find or compile dedicated dictionaries and train a specific classifier for each topic, which would be difficult and expensive beyond our purposes. Finally, a topic-generic approach is necessary for our system to be lightweight and applicable to any user query.

Using SentiStrength we obtain a positive and a negative score for each input text, which we combine into a *polarity* score (Kucuktunc et al. 2012), measuring the inclination towards positive or negative sentiment. By default, the tool computes document-level sentiment scores. Since a document can mention many different entities, and the sentiment around them may vary considerably, we compute entity-level scores by considering small windows (20 words) of text around each mention of an entity, and then averaging across all mentions, similarly to our previous work (Bordino et al. 2013a).

#### 4.2.2 Algorithms

We describe six bundling algorithms, which were based on how to exploit some of the above metadata in a different way. Our aim was not to exhaustively explore all possible bundling approaches, but to study how returning bundled entities compares to returning entity lists in promoting explorative search. We experiment with these six bundling approaches to ensure that our results do not depend on the specifics of one approach, as well as gaining understanding of what type of bundles work best in promoting exploratory search.

For each algorithm we provide a textual, intentionally informal description, and a compact pseudo-code that summarizes the most important steps it performs. Table 4 explains the symbols that are common to all algorithms. The specific symbols used to indicate the bundles created by each algorithm vary, using additional superscripts and/or subscripts to refer to the specific metadata used by the algorithm (categories, subtopics, polarity values).

**1. Bundling based on the categories of the query entity.** Given a query entity, our first algorithm, called `query-categories`, performs the baseline random walk in the entity graph, with restart to the query (Sect. 4.1). It then produces up to 3 result bundles, one for each category of the query (each query is assigned 3 categories, as described in Sect. 4.2.1). The bundle of a category is populated by taking from the baseline ranking vector, the top  $n$  entities belonging to the category. This approach may produce overlapping bundles. We consider this reasonable because the categories of many entities naturally overlap.

---

#### Algorithm 1 `query-categories(q)`

---

- 1: Compute  $\mathcal{R}_{\{q\}}(v)$
  - 2:  $\forall e \in \mathcal{C}(q)$  compute  $\mathcal{B}_1^c(q), |\mathcal{B}_1^c(q)| \leq n$ , containing
  - 3: the  $n$  entities with  $\max \mathcal{R}_{\{q\}}(e)$  and  $c \in \mathcal{C}(e)$ ,  $\forall e \in \mathcal{B}_1^c(q)$
- 

*Example* *Indigenous Australians* has 3 categories: (1) “Society and Culture/Cultures and Groups” which includes *National Sorry Day*—an annual event held in Australia to

<sup>13</sup> <http://sentistrength.wlv.ac.uk/documentation/SentiStrengthChapter.pdf>

**Table 4** Glossary of symbols used in pseudo-code

|                          |  |
|--------------------------|--|
| $\mathcal{R}_{\{q\}}(v)$ | The stationary distribution of the random walk with restart to set of nodes $\{q\}\{q\}$ |
| $\mathbf{C}$             | The whole set of categories used in the system   |
| $\mathcal{C}(q)$         | The categories assigned to entity $q$  |
| $\mathcal{B}_i(q)$       | The bundles created by Algorithm $i$ for query $q$                                       |
| $\mathcal{S}(q)$         | The set of entities specializing entity $q$  |
| $pol(e)$                 | The polarity score of entity $e$   |

commemorate the mistreatment of indigenous people, (2) “Arts and Humanities/History” with historical events like massacres and passage of laws, and (3) “Travel/Australia”, which includes the Jamison Valley.

**2. Bundling based on categories shared by the result entities.** This algorithm, dubbed `result-categories`, computes the baseline ranking vector to retrieve the top  $n$  results for a query, and then groups those based on common categories. The algorithm first attempts to build the most unlikely bundles, those consisting of entities that have exactly the same 3 categories. The requirement is then relaxed to 2 and to 1 category.

---

**Algorithm 2** `result-categories(q)`

---

- 1: Compute  $\mathcal{R}_{\{q\}}(v)$
  - 2: For  $(j = 3 \text{ to } 1)$  compute  $\{\mathcal{B}_{2_i}^j(q)\}$ , s.t.
  - 3:  $\exists \mathcal{C}_i \subseteq \mathbf{C}$ , s.t.  $|\mathcal{C}_i| = j$ , and  $|\mathcal{B}_{2_i}^j(q)| \leq n$  contains
  - 4: the  $n$  entities with  $\max \mathcal{R}_{\{q\}}(e)$  and  $\mathcal{C}_i \subseteq \mathcal{C}(e) \forall e \in \mathcal{B}_{2_i}^j(q)$
- 

*Example* The results for *Porcelain* are divided into (1) “Arts and Humanities” which includes *Glaze defects* (flaws in the surface quality of ceramic glaze), including *Toothpaste* and *CEREC* (a method for creating dental restorations), and (3) “Travel” which includes *Royal Copenhagen*, the Royal Porcelain Factory.

**3. Bundling based on entity specializations.** This algorithm, dubbed `specializations`, takes the top 3 entity specializations of the query and creates a bundle for each. Differently from the other cases, this method modifies the original baseline. For each specialization the random walk in the graph is performed with restart to the query *and* the specialized entity, to capture results that are related to both. The bundle of a specialization is then populated with the top  $n$  entities from the corresponding ranking vector.

---

**Algorithm 3** `specializations(q)`

---

- 1:  $\forall q_j \in \mathcal{S}(q)$  :
  - 2: Compute  $\mathcal{R}_{\{q, q_j\}}(v)$
  - 3: Build  $\mathcal{B}_3(q, q_j)$ ,  $|\mathcal{B}_3(q, q_j)| \leq n$ , containing
  - 4: the  $n$  entities with  $\max \mathcal{R}_{\{q, q_j\}}(e)$
- 

*Example* *Major depressive disorder* is specialized with (1) “Symptom” which includes *Insomnia* and *Hallucination*, (2) “Therapy” including *Psychotherapy* and *Psychoanalysis*, and (3) “Pharmaceutical drug” like *Antidepressant* and *Buprenorphine*.

**4. Bundling based on polarity.** This algorithm, dubbed `polarity`, takes the top  $n$  entities from the baseline ranking, and divides them into 3 bundles based on polarity scores: a *positive* bundle (polarity = 1), a *negative* bundle (polarity = -1), and a *neutral* one (polarity = 0).

---

**Algorithm 4** `polarity(q)`


---

- 1: Compute  $\mathcal{R}_{\{q\}}(v)$
  - 2: For  $j \in \{-1, 0, 1\}$  compute  $\mathcal{B}_4^j(q), |\mathcal{B}_4^j(q)| \leq n$ , containing
  - 3: the  $n$  entities with  $\max \mathcal{R}_{\{q\}}(e)$  and  $pol(e) = j \forall e \in \mathcal{B}_4^j(q)$
- 

*Example* The results for *Libya* are divided into (1) “Positive”, with other countries like *Morocco* and *Egypt*, (2) “Negative”, including *War on Terror* and *Muammar Al-Gaddafi*, with (3) “Neutral” remaining empty.

**5. Categories then polarity.** The `categories-then-polarity` algorithm first creates topical bundles by using `query-categories`. Next, it divides each topical bundle into 3 polarity bundles.

---

**Algorithm 5** `categories-then-polarity(q)`


---

- 1: Compute  $\mathcal{B}_1(q) = \text{query-categories}(q)$
  - 2:  $\forall \mathcal{B}_{1i}(q) \in \mathcal{B}_1(q), \forall j \in \{-1, 0, 1\}$  :
  - 3: Compute  $\mathcal{B}_{5i}^j(q) \subset \mathcal{B}_{1i}(q)$  s.t.  $pol(e) = j \forall e \in \mathcal{B}_{5i}^j(q)$
- 

**6. Polarity then categories.** The `polarity-then-categories` algorithm first creates 3 polarity bundles using `polarity`, then it further divides each polarity bundle into topical bundles, based on the categories of the query as in `query-categories`.

---

**Algorithm 6** `polarity-then-categories(q)`


---

- 1: Compute  $\mathcal{B}_4(q) = \text{polarity}(q)$
  - 2:  $\forall \mathcal{B}_{4i}(q) \in \mathcal{B}_4(q), \forall c \in \mathcal{C}(q)$  :
  - 3: Compute  $\mathcal{B}_{6i}^c(q) \subset \mathcal{B}_{4i}(q)$  s.t.  $c \in \mathcal{C}(e) \forall e \in \mathcal{B}_{6i}^c(q)$
- 

We did not combine `result-categories` or `specializations` with `polarity` because preliminary experiments showed that `query-categories` was clearly outperforming them, as confirmed by the study in Sect. 8.

## 5 Bundle characterization

Applying the above algorithms to our 150 test queries, we now examine their behavior in terms of similarity of their result sets, cohesiveness and diversity of the bundles created, and the tendency to select results with high or low baseline rank, indicating how far each “explores” the initial result set.

## 5.1 Result set similarity

We first compare the six algorithms in terms of the similarity of the results they produce. Each method uses different semantic information, so we expect the bundles built by different algorithms to be conceptually different.

For each algorithm and for each query, we consider the union of the entities in the bundles produced for the query. We compare each algorithm to all the other algorithms in terms of the distribution of the Jaccard similarity of the result sets returned for the test queries. Table 5 reports the average and maximum Jaccard similarity obtained for each pair of algorithms. The algorithms are numbered following the order in Sect. 4.2.

The algorithms tend to pick different results for a query, which was expected since they use different metadata and require different constraints. Experimenting with different bundles is important to evaluate how users perceive bundled results compared to typical item-based results, as some bundles may lead to a better experience than others. The average similarity between the result sets of two different methods is very low—it never exceeds the value of 0.3. The third (*specializations*) and the fifth (*categories-then-polarity*) algorithm produce the most different results: the comparisons involving them achieve the lowest similarity values. Indeed, *specializations* is the most different algorithm by design—it relies on a random walk different from the baseline, while the latter requires a more complex constraint.

## 5.2 Cohesiveness and separation

After verifying that the bundling algorithms pick different results for a query, we also investigate whether they create cohesive and well-separated clusters with respect to the textual similarity measure used to build the network. We remark that all of the six bundling algorithms group entity results by employing additional semantic information, i.e., topic (super-categories or specializations), sentiment, or a combination of both, which is still derived from the data but not captured by the sole entity network. This enriched information is the key ingredient to bundle results, thus we do not expect perfect clusters when looking at the sole textual similarity of entity results.

Given that our baseline algorithm simply builds a ranked list of entities and we do not have a ground-truth clustering to compare to, we consider various unsupervised cluster evaluation measures, which evaluate the goodness of a clustering structure. Unsupervised measures of validity are often further divided into two classes: measures of cluster *cohesion* (or compactness, tightness), which determine how closely related the objects in a cluster are, and measures of cluster *separation* (diversity, isolation), which determine how distinct or well-separated a cluster is from other clusters.

**Table 5** Jaccard similarity of the result sets produced by the six bundling algorithms

| Algs | Avg   | Max   | Algs | Avg   | Max   | Algs | Avg   | Max   |
|------|-------|-------|------|-------|-------|------|-------|-------|
| 1/2  | 0.131 | 0.476 | 1/3  | 0.033 | 0.185 | 1/4  | 0.170 | 0.500 |
| 1/5  | 0.049 | 0.200 | 1/6  | 0.297 | 0.889 | 2/3  | 0.032 | 0.258 |
| 2/4  | 0.298 | 0.684 | 2/5  | 0.055 | 0.182 | 2/6  | 0.181 | 0.476 |
| 3/4  | 0.074 | 0.250 | 3/5  | 0.015 | 0.088 | 3/6  | 0.042 | 0.179 |
| 4/5  | 0.026 | 0.160 | 4/6  | 0.103 | 0.412 | 5/6  | 0.137 | 0.833 |

**Table 6** Internal cluster validation measures

| Alg | Cohesion |       |       | Separation |       |       | Dunn      | Silhouette |
|-----|----------|-------|-------|------------|-------|-------|-----------|------------|
|     | Median   | Avg   | Max   | Median     | Avg   | Max   |           |            |
| 1   | 0.528    | 0.365 | 0.981 | 0.000      | 0.234 | 0.988 | 0.0001535 | 0.009308   |
| 2   | 0.537    | 0.397 | 0.988 | 0.505      | 0.310 | 0.980 | 0.0001483 | −0.023090  |
| 3   | 0.000    | 0.231 | 0.977 | 0.000      | 0.081 | 0.974 | 0.0002987 | 0.014180   |
| 4   | 0.513    | 0.320 | 0.972 | 0.000      | 0.183 | 0.974 | 0.0003536 | 0.047250   |
| 5   | 0.515    | 0.332 | 0.967 | 0.000      | 0.209 | 0.967 | 0.0004803 | 0.010740   |
| 6   | 0.534    | 0.381 | 0.981 | 0.000      | 0.271 | 0.976 | 0.0003461 | 0.009791   |

Our aim is to verify whether the bundles created by the algorithms correspond to internally cohesive and externally well-separated portions of the entity network. Thus we adopt simple graph-based notions of cluster validity (Tan et al. 2005): for each algorithm and for each query, we compute cohesion as the sum of the similarity weights of the arcs connecting any pair of entities in the same bundle, and separation as the sum of the similarities of the arc connecting any pair of distinct entities output by the algorithm in different bundles. The first two columns in Table 6 report the distribution of cohesive and separation obtained, showing the median, average and maximum intra- and inter-bundle similarity achieved by each algorithm over the 150 test queries (the algorithms are numbered following Sect. 4.2).

The six algorithms exhibit a similar behavior: for all of them we observe a low average value of separation, ranging from 0.081 and 0.31, indicating that results in different bundles are not very related to each other, which is a desideratum. However, the average value of cohesion or internal similarity is also low (ranging from 0.231 to 0.397), even if it is always higher than the corresponding inter-bundle similarity. The lowest similarity values are achieved by the third algorithm (*specializations*), which picks results that are farther away from the query.

Other measures of cluster validity (Liu et al. 2010) combine the two aspects of cohesion and separation: for example, the method of Silhouette (Rousseeuw 1987) coefficients measures the difference of between- and within- cluster distances. Dunn's index (Dunn 1974) computes the ratio between cluster separation, measured as minimum distance between clusters, and cohesion measured as maximum distance in between data points of clusters. For each bundling algorithm and for each query we computed Silhouette and Dunn's index using cosine distance as the distance measure. The rightmost columns in Table 6 reports the average results obtained over the 150 test queries. Once again, in all cases we obtain very low values, indicating that the bundling methods do not seem to create good clusters according to cosine distance.

In retrospect, the fact that for all algorithms, the results in the same bundle do not strongly relate to each other—for what concerns the basic similarity measure used in the system, i.e. the syntactic similarity of the context where the entities appear—indicates that the additional metadata exploited by each algorithm (categories, subtopics, polarity or a combination of these) conveys important semantic information about the entities, which is not captured by the network alone, confirming and motivating the usage of such metadata for creating meaningful organizations of search results.



**Table 7** Baseline rank of results in bundles

| Algorithm | Median | Avg     | Max     |
|-----------|--------|---------|---------|
| 1         | 14     | 42      | 2887    |
| 2         | 13     | 4462    | 896,800 |
| 3         | 1712   | 169,700 | 896,800 |
| 4         | 12     | 12      | 28      |
| 5         | 193    | 487     | 25,080  |
| 6         | 29     | 30      | 83      |

### 5.3 Relation with the baseline rank

Finally, we analyze the extent to which the algorithms utilize the initial results, operationalized by the rank of the items they select from the original result set. Table 7 shows the median, average and maximum baseline rank for the results of each algorithm, aggregating over all bundles and all queries. Here, lower rank identifies results higher up in the original result set.

As expected, the algorithms which use the simplest constraints pick more results with high rank—these are the first (*query-categories*) and the fourth (*polarity*). The former picks results in a category, and the latter with a given polarity value. The other algorithms must go further down in the rank to find entities that satisfy more complex constraints, such as sharing up to three categories with all results in the same bundle (*result-categories*), or a combination of topical and polarity constraints (the hybrid algorithms). Not surprisingly, the algorithm selecting the entities lowest in the rank is the third (*specializations*), which is based on a different random walk.

Our analysis suggests that the bundling algorithms produce different results, employing different metadata to discover relations among entities that are not captured in the sole entity network. Simpler algorithms may be preferable because complex constraints might excessively hurt the relevance of results. This will be confirmed by the studies in Sects. 7 and 8.

## 6 Implementation

The basic implementation of our framework is described in (Laere et al. 2014). Our tool, dubbed DEESSE, served as a demonstrator of the concept of serendipitous entity search in the EU project LiMoSINE.<sup>14</sup> In the present paper, aimed at investigating whether bundling search results can lead to an improved explorative search experience, we leverage our tool to conduct large-scale user studies with workers recruited through Mechanical Turk.

Before presenting the extensive evaluation of our bundling approaches, we briefly recall the main features of the implemented tool, and explain how users can interact with it to explore enriched entity networks extracted from Yahoo Answers (or other social media of interest).

### 6.1 Architecture

The version employed in our user studies extends the original tool presented in (Laere et al. 2014), providing support for the Italian language (in addition to the original English

<sup>14</sup> [www.limosine-project.eu](http://www.limosine-project.eu).

and Spanish) and multi/cross-lingual search,<sup>15</sup> and implementing all the six bundling algorithms described in this paper. The architecture of such module consists of a back-end and a front-end parts. The back-end works offline, extracting an enriched entity network from the considered dataset, and precomputing bundled result sets for every entity. The core components of the back-end are the modules are: entity-network extractor, entity feature extractor, baseline ranking algorithm, algorithms for bundled retrieval.

Although we use a distributed parallel implementation to perform random-walk computations on large-scale data efficiently, our basic retrieval algorithm requires a temporal computational cost of minutes to obtain results for a query entity. This cost makes running the ranking module at query time prohibitive. To make our solution viable, we perform the computation offline. To avoid storing the full stationary distribution of every node, we also run the bundling algorithms offline, and only store in an index the resulting bundles obtained for each entity. When the number of query entities grows, the computation becomes even more expensive. These constraints make that a daily update of the data is not worth the minor improvement in query behaviour for the user. Having a slower process implies that we are not able to serve extremely recent time-sensitive queries. However, those queries are not a critical use case in our scenario (exploration of user-generated content such as Yahoo Answers data), thus we consider the above limitation acceptable.

The front-end receives the query submitted by the user and sends a request to the back-end to retrieve the corresponding results. The technologies used for the front-end consist of a combination of a MySQL database, PHP, CSS, HTML and Javascript. D3.js<sup>16</sup> is being used to retrieve JSON formatted data and manipulate it for display, whereas the Bootstrap<sup>17</sup> framework is used to style the web interface. When a query comes from the front-end, the resulting (pre-computed) bundles are retrieved from the index. Complimentary metadata is provided by the entity feature extractor (e.g., sentiment) or fetched from external sources (abstract or image urls of Wikipedia page and top-rated Yahoo Answers question/answer pair).

## 6.2 Interaction

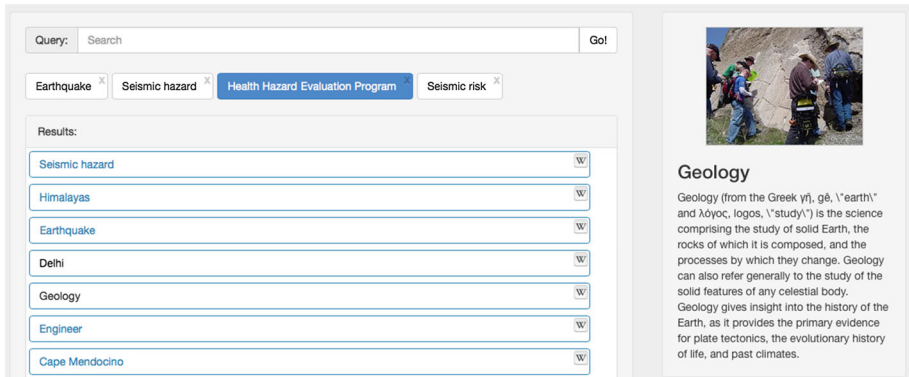
Whenever a user interacts with DEESSE, the search results for the query entity are presented in the central panel of the web user interface. The results are either presented in a ranked list or grouped into bundles. Figure 1 shows the ranked list of results returned by the baseline algorithm for the query “Health hazard evaluation program”. In the case of bundled results, for each of the entities in a bundle, an illustration of sentiment polarity is provided (if available), along with a link to the Wikipedia page of that entity. A click on an entity result will initiate a search with this entity as query. Hovering an entity in the result list will trigger the retrieval of any available metadata from Wikipedia (thumbnail picture and Wikipedia abstract) and Yahoo Answers (top rated question and answer mentioning the entity). Multiple searches can be carried out, and buttons will appear under the search bar to keep track of them. Clicking one of the previous searches will again show the results for that specific search, whereas clicking the close button in the top-right corner will remove the results for that query entity.

Figure 2 shows the results returned by the `query-categories` algorithm for the query “NASA”. The closer entities retrieved from the entity network are grouped into

<sup>15</sup> We do not discuss this aspect of the tool as it is outside the scope of this paper.

<sup>16</sup> <http://d3js.org/>.

<sup>17</sup> <http://getbootstrap.com/>.



**Fig. 1** Ranked-list results for the query “Health hazard evaluation program”

three bundles, based on the top three categories associated with those entities in the Yahoo Answers dataset. Given that each entity is associated with three categories, the bundles may overlap. In this example, the first category associated with NASA is *Astronomy and Space*, including results such as *Space Shuttle*, *Spaceflight Hubble Space Telescope*, *Moon landing*, entities such as the *International Space Station* or the *European Space Agency*, planets such as *Jupiter*, *Pluto*, *Venus*. The second category is *Politics and Government/Politics*, showing that people who talk about NASA in Yahoo Answers often discuss its connections with CIA activities or discuss conspiracy theories and the possibility that some of the NASA programs were faked. In the third category, *Society and Culture/Religion and Spirituality* we can find references to mystery and strange beliefs, conspiracy theories, mythology and extraterrestrial life.

In this paper, we use DEESSE to conduct three user studies, described in the upcoming sections. The first two studies involve annotation tasks comparing a bundling algorithm to the baseline retrieval algorithm and two another bundling algorithm, respectively. The comparison tasks that are shown to annotators are built reusing DEESSE interfaces. In the third study, annotators use DEESSE to simulate a more realistic and interactive search activity.

## 7 To bundle or not to bundle

In the rest of this paper, we describe three studies and their results performed using Amazon Mechanical Turk<sup>18</sup> (MT), summarized in Table 8. First, we compare our bundling algorithms to the ranked-list baseline, then we find the best approach, and finally we test it in a more realistic setting of an interactive search simulation. In the first two studies, we focus on quantitative evaluation, collecting thousands of responses, whereas in the last we administer an extensive survey for a more qualitative feedback. In this section, our first study, we investigate whether users prefer composite answers over standard ranked lists.

<sup>18</sup> <https://www.mturk.com/mturk/> .



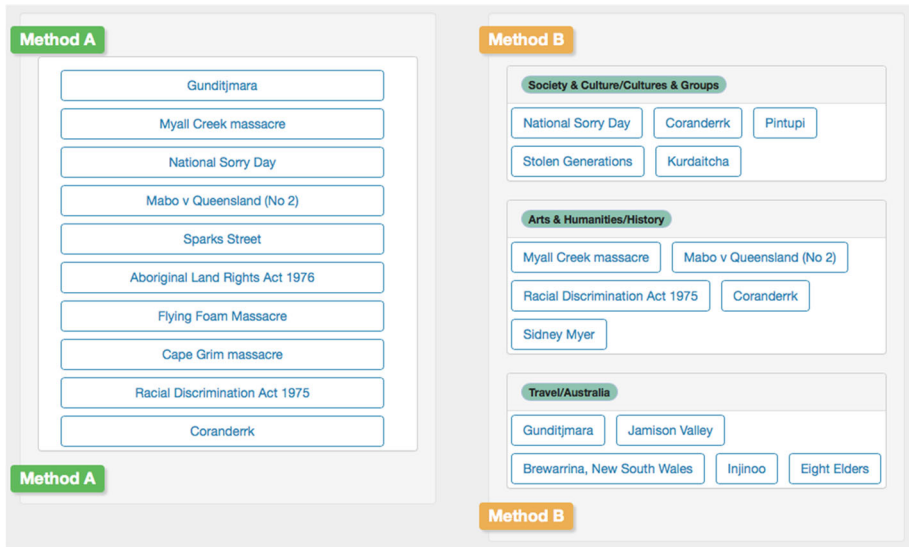
Fig. 2 Bundled search results for the query “NASA”

Table 8 Summary of the studies

| # | Goal                          | Kind        | Responses |
|---|-------------------------------|-------------|-----------|
| 1 | Bundles versus ranked lists   | Labeling    | 2002      |
| 2 | Best bundling approach        | Labeling    | 996       |
| 3 | Explorative search experience | Interactive | 298       |

### 7.1 Task design

As mentioned in Sect. 4.1, we use 150 entities from the most searched queries in 2010/2011 from Google Zeitgeist. We design annotation tasks to compare the standard ranked list produced by the baseline method to the bundled result set returned by each of our six bundling methods. For each test query we build a comparison task for each of our six bundling algorithm. Each task shows the user a ranked list of the top 15 results returned by the baseline retrieval algorithm, and the top three bundles returned by one bundling algorithm, each containing a maximum of five entities. The two methods under comparison in a task are anonymized, and simply labeled as “Method A” and “Method B”. Both query and results (either in the ranked list or in a bundle) are presented in the form of Wikipedia entities, with a link that can be navigated bringing the user to the corresponding Wikipedia page. The results in the ranked list are sorted by decreasing order or relevance (as returned by the baseline algorithm). The entities returned by a bundling algorithm in return to a query, are presented grouped in bundles with are labeled with the relevant information



**Fig. 3** Illustration of the way the comparison task between the baseline (ranked list) and the bundling strategy query-categories was presented to the annotators, showing the results for the query entity “Indigenous Australians”

characterizing a bundle (a category, an entity specialization or subtopic, polarity, or a combination of these).

As an example, Fig. 3 illustrates the comparison task between the baseline (ranked list) and the bundling algorithm query-categories that was presented to annotators in the case of the query entity “Indigenous Australians”, which has 3 categories: (1) “Society and Culture/Cultures and Groups” that includes *National Sorry Day*—an annual event held in Australia to commemorate the mistreatment of indigenous people, (2) “Arts and Humanities/History” with historical events like massacres and passage of laws, and (3) “Travel/Australia”, which includes the Jamison Valley.

The user is provided with the following instructions: “You are assigned a query and two alternative sets of results returned as answers to a query. Look at the query and at the various results. In case you need more information about the query or any of the results, click on it and you will be directed to the corresponding Wikipedia page. Once you have got a good sense of the query and the two result sets, please answer the questions at the end of the task. You will be asked to compare the two result sets and to indicate which one you prefer. You will be asked your Worker ID to login to the survey, and will be presented a code upon the completion of the survey. You can take the survey multiple times until you have annotated all tasks”.

Figure 4 shows the preview of a HIT shown to an annotator: the HIT, dubbed “Comparing result sets”, shows the text reported above, explaining that the task is a simulated search for the query, for which two alternative sets of results are proposed. The HIT then shows a link to the survey, which we hosted on a server of the LiMoSINe project. The user is invited to navigate the links to make sure she has a reasonable idea of the topics and concepts expressed by each entity (query or result), before answering a number of questions. Upon completion of a task we provide the annotator with a code that she must provide back on MT to claim payment.

HIT Preview

**Comparing result sets**

You are assigned a query and two alternative sets of results returned as answers to a query. Look at the query and at the various results. In case you need more information about the query or any of the results, click on it and you will be directed to the corresponding Wikipedia page. Once you have got a good sense of the queries and the two result sets, please answer the questions at the end of the task. You will be asked to compare the two result sets and to indicate which one you prefer.

You will be asked your Worker ID to login to the survey, and will be presented a code upon completion of the survey. You can take the survey multiple times until you have annotated all tasks.

Survey link: [http://deesse.limosine-project.eu/user\\_evaluation/index.php](http://deesse.limosine-project.eu/user_evaluation/index.php)

Provide the survey code here:

Submit

**Fig. 4** Instructions shown to the annotator

We asked the annotators to answer six questions:

- *Q1*: “Are you familiar with the topic of the query?”
- *Q2*: “Which result set is better organized?”
- *Q3*: “Which result set is better at revealing the most useful/interesting results?”
- *Q4*: “Which result set helped you to understand the search results?”
- *Q5*: “Which result set is easier to get a good sense of the range of alternatives? (Diversity of the result)”
- *Q6*: “Which result set helped you to find new topics related to the query?”

The answer to *Q1* consisted of a value ranging from 1 (not familiar) to 5 (very familiar), using a Likert scale, while the other questions had a set of options defined by “*Both bad*”, “*Method A*”, “*Method B*” or “*Both good*”. We required at least three distinct annotators for each of the 150 tasks.

## 7.2 Task serving

We recruited MT workers living in the United States, having HIT (Human Intelligence Tasks) Approval Rates of over 95 % and at least 1000 approved HITs. To filter malicious or fraudulent submissions we created a set of 18 *gold* tasks, for which we provided ranked-list and bundling results. Half of these tasks consisted of real results, for which a human assessor can immediately observe a clear preference for one of the methods. The other half of the tasks consisted of artificially generated (rubbish) results for one of the methods, clearly pointing towards a positive evaluation of only one method. Using the known outcome for these tasks, we assessed the quality of the responses by looking at the answers provided to questions *Q2* and *Q3*.

Workers were served with 11 annotations tasks, consisting of 8 real and 3 gold tasks, served in a random order. The tasks were sampled at random from the missing annotations for our study. The answers to each of the questions were randomized. During the task execution, we tracked for each participant the time spent on each individual task, and whether the user answered the gold questions correctly. Each user received six golden questions per annotation session (*Q2* and *Q3* for the 3 gold tasks), and we retained as *trusted* workers only those who answered correctly more than half of the received gold questions. We filtered the annotations of untrusted workers, yielding 124 entities shared over all of the methods for our experiment that received at least one annotation. However we attained annotations by 3 or more distinct workers for the majority of the entities.

Using the outlined methodology, we gathered (after filtering) 2002 valid annotations by 142 distinct trusted workers for all the required comparisons. The workers spent on average 60.9 s on a single annotation task. The average familiarity was 3.5 on a 5-point scale, indicating that the selected queries were known to the MT workers.

### 7.3 Comparison results

Table 9 presents the results for the 124 queries that were annotated, along with label overlap between the participants. We present the 5 questions (*Q2–Q6*) asking the user to express a preference for either the ranked list or the bundled result set produced by the

**Table 9** Comparing each bundling method over baseline ranked list

| Method                   | Question                   | Bundled list | Ranked list | Both good | Both bad | Agree (%) |
|--------------------------|----------------------------|--------------|-------------|-----------|----------|-----------|
| query categories         | <i>Q2</i> <sup>*****</sup> | <b>74</b>    | 21          | 11        | 18       | 73        |
|                          | <i>Q3</i> <sup>*****</sup> | <b>95</b>    | 15          | 5         | 9        | 78        |
|                          | <i>Q4</i> <sup>*****</sup> | <b>74</b>    | 17          | 15        | 18       | 70        |
|                          | <i>Q5</i> <sup>*****</sup> | <b>73</b>    | 20          | 10        | 21       | 73        |
|                          | <i>Q6</i> <sup>*****</sup> | <b>68</b>    | 20          | 16        | 20       | 70        |
| result categories        | <i>Q2</i>                  | <b>48</b>    | 42          | 21        | 13       | 69        |
|                          | <i>Q3</i> <sup>*****</sup> | <b>66</b>    | 25          | 21        | 12       | 68        |
|                          | <i>Q4</i>                  | <b>52</b>    | 34          | 23        | 15       | 66        |
|                          | <i>Q5</i>                  | <b>52</b>    | 34          | 23        | 15       | 72        |
|                          | <i>Q6</i>                  | <b>48</b>    | 34          | 26        | 16       | 65        |
| specializ.               | <i>Q2</i>                  | <b>47</b>    | 42          | 14        | 21       | 72        |
|                          | <i>Q3</i> <sup>**</sup>    | <b>64</b>    | 32          | 12        | 16       | 75        |
|                          | <i>Q4</i>                  | <b>51</b>    | 39          | 12        | 22       | 73        |
|                          | <i>Q5</i>                  | <b>48</b>    | 42          | 9         | 25       | 75        |
|                          | <i>Q6</i>                  | <b>45</b>    | 43          | 14        | 22       | 71        |
| polarity                 | <i>Q2</i> <sup>*****</sup> | 23           | <b>62</b>   | 13        | 26       | 76        |
|                          | <i>Q3</i> <sup>***</sup>   | 27           | <b>59</b>   | 20        | 18       | 66        |
|                          | <i>Q4</i> <sup>**</sup>    | 29           | <b>59</b>   | 14        | 22       | 75        |
|                          | <i>Q5</i> <sup>*****</sup> | 22           | <b>58</b>   | 15        | 29       | 72        |
|                          | <i>Q6</i> <sup>*****</sup> | 20           | <b>60</b>   | 20        | 24       | 68        |
| categories then polarity | <i>Q2</i>                  | <b>49</b>    | 42          | 19        | 14       | 70        |
|                          | <i>Q3</i> <sup>*</sup>     | <b>62</b>    | 39          | 15        | 8        | 73        |
|                          | <i>Q4</i> <sup>*</sup>     | <b>58</b>    | 36          | 17        | 13       | 72        |
|                          | <i>Q5</i>                  | <b>48</b>    | 41          | 14        | 21       | 76        |
|                          | <i>Q6</i>                  | <b>48</b>    | 44          | 15        | 17       | 74        |
| polarity then categories | <i>Q2</i>                  | 38           | <b>52</b>   | 19        | 15       | 71        |
|                          | <i>Q3</i>                  | 44           | <b>48</b>   | 22        | 10       | 69        |
|                          | <i>Q4</i>                  | 40           | <b>49</b>   | 19        | 16       | 68        |
|                          | <i>Q5</i> <sup>**</sup>    | 33           | <b>53</b>   | 20        | 18       | 67        |
|                          | <i>Q6</i>                  | 35           | <b>47</b>   | 25        | 17       | 66        |

*p* < 0.0001\*\*\*\*; *p* < 0.001 \*\*\*; *p* < 0.01 \*\*; *p* < 0.05 \*

considered bundling algorithm. For each question we report the number of times that each of the possible answers (ranked list, bundling, or both good or both bad) was selected by the majority of the annotators. In each row we mark in bold the answer that won the highest number of comparisons for the given question. Statistical significance, computed using the Wilcoxon signed-rank test, is reported—when found—for each question.

The results of this study clearly indicate a preference for *query-categories*, which creates bundles using the super-topics of the query entity. The bundled result set is preferred over the ranked list in the large majority of comparisons (ranging from 55 to 77 %). The agreement (label overlap) ranges from 70 to 78 %. The results obtained for *query-categories* are statistically significant with  $p < 10^{-4}$ .

The other two algorithms for topical bundling, i.e., *result-categories*, which groups results based on common super-topics, and *specializations* that creates a bundle for each sub-topic of the query entity, are also always beating the baseline ranked list for all questions, but they never perform as well as *query-categories*. *Result-categories* won from 39 to 53 % of comparisons, and *specializations* from 36 to 52 %. Both methods were still considered better than the baseline at revealing more interesting and useful results, with results for *Q3* being statistically significant.

When bundles are determined according to sentiment (with *polarity*), the ranked-list results are preferred over the bundles, achieving strong statistical significance for all questions. When combining categories with sentiment (with *categories-then-polarity*), the bundling is preferred over the ranked list, albeit less convincingly. Finally, when combining topic and sentiment the other way around (*polarity-then-categories*), the ranked list performs better.

We can conclude that using bundling methods is useful for improving search results, as long as they are constructed using the right type of metadata, in our case topics, and especially the super-categories of the query entity. This result echoes that of a recent study on manual construction of bundles (Bota et al. 2015), which found that users prefer bundles which are topically cohesive and relevant. Also, as we discuss in Sect. 10, it could be the case that not all topics are equally suited for sentiment metadata, with select few cohesively clustering around different emotions. How bundling helps in a real search task is something we investigate in Sect. 9. Before that, we compare bundles themselves.

## 8 Identifying the best bundling

In this section, we seek to find the best bundling algorithm by comparing the three which performed the best thus far, i.e. the ones based on topic metadata. Our aim is to study how users perceive bundles formed with different approaches, and to identify the best approach, which we further evaluate in a realistic search task in the next section.

### 8.1 Task design

We continued with the same 150 entities considered in the previous study. For each, we extracted bundled result sets with the 3 topical bundling algorithms, generating the top best 3 bundles, with a maximum of 5 entities in each. We generated 3 annotation tasks that compared the possible *combinations* of the three methods: (1) *query-categories* versus *result-categories*, (2) *query-categories* vs. *specializations*, and (3) *result-categories* versus *specializations*.



Each task now shows the user two result sets for a query, built with two different bundling algorithms. Each result set is formed by at most three bundles, and each bundle contains a maximum of five entities. The two bundling methods under comparison in a task are anonymized, and simply labeled as “Method A” and “Method B”. Both query and results in the bundles are presented in the form of Wikipedia entities, with a link that can be navigated bringing the user to the corresponding Wikipedia page. Each bundle is labeled with the relevant information characterizing it (a category associated with the query in the case of *query-category*, a category shared among the results included in the bundle in the case of *result-categories*, and a query specialization or subtopic in the case of *specializations*). Figure 5 illustrates an example of comparison task for the query entity “Indigenous Australians”.

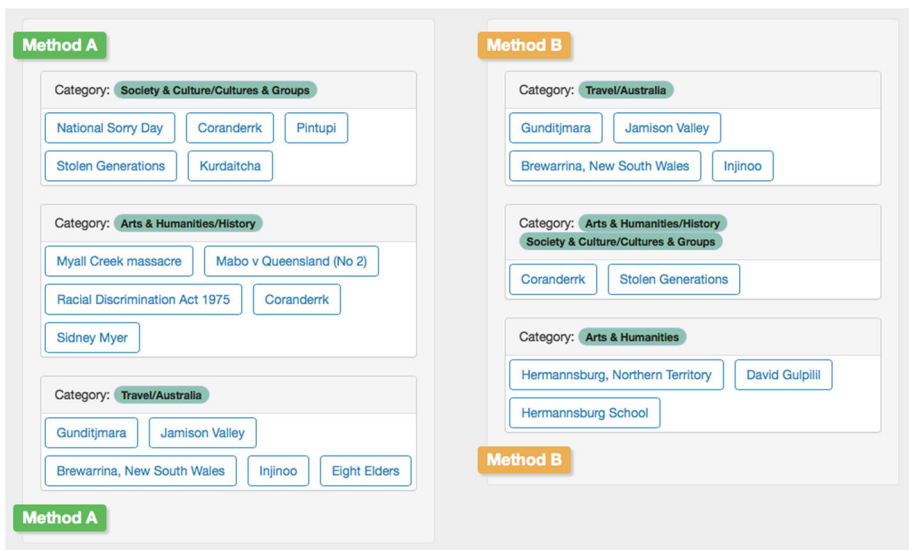
The user is provided with the same instructions that were used in the previous experiment: “You are assigned a query and two alternative sets of results returned as answers to a query. Look at the query and at the various results. In case you need more information about the query or any of the results, click on it and you will be directed to the corresponding Wikipedia page. Once you have got a good sense of the queries and the two result sets, please answer the questions at the end of the task. You will be asked to compare the two result sets and to indicate which one you prefer. You will be asked your Worker ID to login to the survey, and will be presented a code upon the completion of the survey. You can take the survey multiple times until you have annotated all tasks”.

Figure 6 shows the preview of a HIT shown to an annotator. We used the same mechanism as before to direct the user to a survey on our servers, and provide her with a code that she could use to claim payment on MT after completing the annotation task.

For our test queries we built 450 annotation tasks, defined by an *entity-combination*-pair, requiring an annotator to provide answers to the following questions:

Q1: “Are you familiar with the topic of the query?”

Q2: “Which result set is better organized?”



**Fig. 5** Illustration of the way the bundling comparison task was presented to the annotators, showing the results for the query entity “Indigenous Australians”

**Comparing result sets**

You are assigned a query and two alternative sets of results returned as answers to a query. Look at the query and at the various results. In case you need more information about the query or any of the results, click on it and you will be directed to the corresponding Wikipedia page. Once you have got a good sense of the queries and the two result sets, please answer the questions at the end of the task. You will be asked to compare the two result sets and to indicate which one you prefer.

You will be asked your Worker ID to login to the survey, and will be presented a code upon completion of the survey. You can take the survey multiple times until you have annotated all tasks.

Survey link: <http://deesse.limosine-project.eu/study2/index.php>

Provide the survey code here:

**Fig. 6** Instructions shown to the annotator

*Q3*: “Which result set is better at revealing the most useful/interesting results?”

*Q4*: “Which grouping is easier to get a good sense of the range of alternatives?  
(Diversity of the result)”

*Q5*: “Which grouping helped you to find new topics related to the query?”

## 8.2 Task serving

We again ran an external questionnaire on MT, using the same set of qualifications as before: we required participants to live in the United States, to have at least 1000 approved HITS and 95 % approval rate. To ensure the quality of the annotations, we created a separate gold set of 20 tasks that consisted of selected entities and a pre-determined comparison of two methods for each of them. As we did in the previous study, in each comparison we showed real results for one method, and rubbish results for the other, so that an annotator could only possibly prefer one of the two methods. In this set of golden comparison tasks, the answers to questions *Q2* and *Q3* were manually labelled and could thus be used for assessing the quality of the annotations. This time, workers were served with 10 annotation tasks, consisting of 8 real and 2 gold tasks, served in a random order. We sampled our missing annotations in a similar way as described in Section 7, and used the same principles of randomization of the answers and metrics for tracking user behavior. Similarly to the previous results, we had to drop some entities that did not receive enough annotations due to filtering of untrusted workers, resulting in 141 entities shared over all comparisons in the experiment.

We gathered 1404 annotations by 165 distinct workers, of which 9 took the questionnaire multiple times. Filtering the untrusted workers left us with 996 valid annotations from 121 unique workers. We did not explicitly forbid participants to take part in multiple experiments, however a large majority of the workers took part in only one study. Only 34 workers participated in both first and second experiments. Similarly to previous study, the average familiarity was 3.496 on a 5-point scale, and the workers spent on average 56.36 s on a single annotation task.

The results in Table 10 identify a clear winner among the 3 topical bundling methods: `query-categories` won the largest fraction of comparisons with both other methods, and for all questions. Observe that `query-categories` was also the winning method in the previous study, where we found that users preferred the bundled result sets returned by this method over the ranked list in the largest fraction of comparisons (ranging from 55 to 77 %). In this experiment, in the majority of the cases the MT workers selected this bundling method as the one providing a better organization of the results, more useful and interesting information, a better coverage of the various aspects of the query and more

**Table 10** Comparing topic-based bundling methods

| Method ( <i>A</i> vs. <i>B</i> )       | Question    | Method <i>A</i> | Method <i>B</i> | Both good | Both bad | Agree |
|--|-------------|-----------------|-----------------|-----------|----------|-------|
| query-categories vs. result-categories | $Q2^{****}$ | 72              | 32              | 28        | 9        | 70    |
|  | $Q3^{****}$ | 73              | 32              | 27        | 9        | 73    |
|  | $Q4^{***}$  | 70              | 31              | 29        | 11       | 70    |
|  | $Q5^{***}$  | 69              | 33              | 32        | 7        | 66    |
| query-categories vs. specializations   | $Q2^{****}$ | 86              | 24              | 24        | 7        | 73    |
|  | $Q3^{****}$ | 90              | 21              | 24        | 6        | 73    |
|  | $Q4^{****}$ | 86              | 26              | 24        | 5        | 73    |
|  | $Q5^{****}$ | 80              | 21              | 34        | 6        | 66    |
| result-categories vs. specializations  | $Q2^{**}$   | 67              | 40              | 19        | 15       | 73    |
|  | $Q3^*$      | 71              | 46              | 15        | 9        | 76    |
|  | $Q4$        | 62              | 42              | 20        | 17       | 73    |
|  | $Q5$        | 62              | 51              | 15        | 13       | 75    |

$p < 0.0001^{****}$ ;  $p < 0.001^{***}$ ;  $p < 0.01^{**}$ ;  $p < 0.05^*$

novelty in the results. The difference between the results obtained for *query-categories* and the other two methods was always statistically significant, with  $p < 10^{-4}$  for most of the questions. Like in previous case, we computed Krippendorff’s alpha coefficient, yielding a median score of 0.16, and a maximum of 0.20.

The second best method was *result-categories*, which won a higher fraction of comparisons with *query-categories*, and was preferred in the majority of cases when compared to *specializations*, although statistical significance in the latter case was only achieved for  $Q2$  and  $Q3$ .

These results, from both the present and the previous sections, suggest that a general, but simple and intuitive topical categorization, such as the one provided by Yahoo Answers, provide a good basis to build coherent and meaningful bundles. The bundles attempting to cover more specific aspects and sub-topics of the query entities, however, were not as much appreciated by the MT workers.

## 9 Explorative search experience

Thus far, the labeling tasks involved overall preference of the participants. We now conduct a third crowdsourced study, using Mechanical Turk. Inspired by the methodology of Borlund and Ingwersen (1997), we create tasks simulating real search scenarios, allowing users to express their views through several questions related to their search experience.

We continue to use Mechanical Turk, as it has been used to perform non-trivial experiments widely. Not only are the workers asked to fill surveys as in (Evans and Chi 2008) and (Capra et al. 2011), they have been widely recruited for complex search tasks such as in (Held and Cress 2009) and (Lagun and Agichtein 2011). Also, Mechanical Turk is an alternative “outside the traditional knowledge worker and student populations” (Capra et al. 2011), making available a much larger and diverse population than what can be found in a university classroom or by other affordable convenience samples.

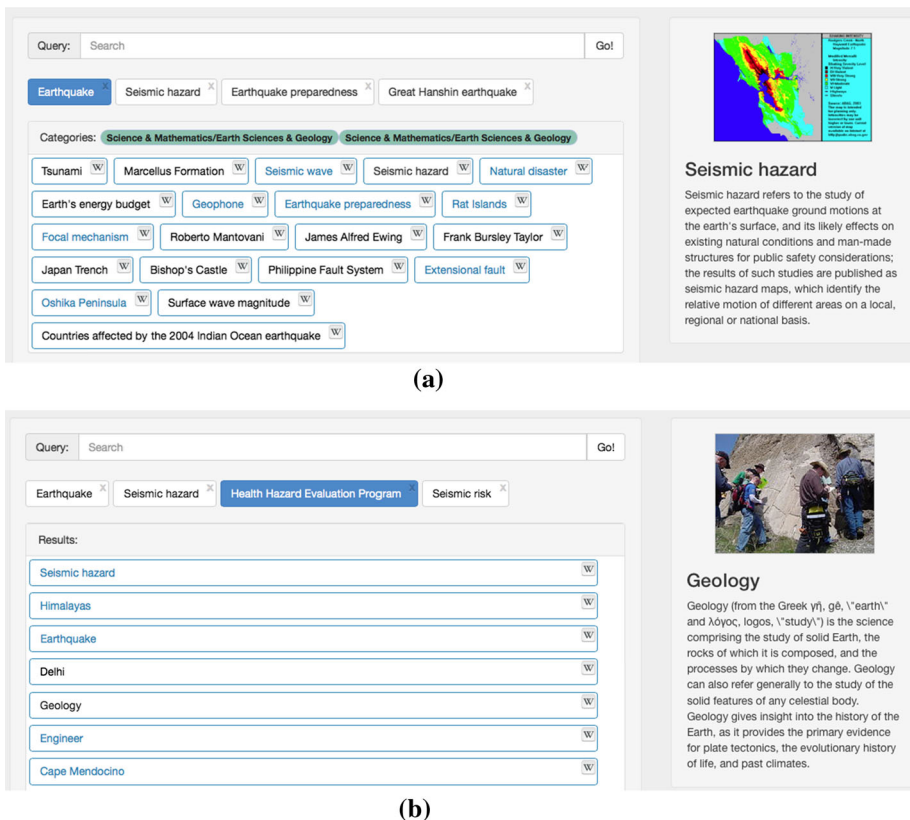
Although it is impossible to customize the tasks to each participant, as suggested by Borlund, we use well-known entities and settings which are likely to be familiar to most people.

We compose 30 simulated explorative search tasks (referred simply to as tasks) for a subset of our 150 query entities. Following Wildemuth and Freund (2012), we ensured that the tasks have learning and investigation as goals, but vary in specificity of the problem. The composition of the tasks was driven by the interfaces themselves, to assure that the tasks are accomplishable, and would not result in frustration.

Examples of tasks for *Earthquake* and *Microfinance* are:

- Assume that you are taking a high-school geology course. You want information about earthquakes and other natural disasters which may occur simultaneously.
- Recently, you became interested in microfinance as an alternative service in small-business territory. Thus, you want to search for the microfinance institutions in the USA.

We use the tasks to compare the experiences users have with the proposed bundling algorithm, as compared to the ranked list. We choose, as our bundling approach, *query-categories*, which was performing the best, as demonstrated in the previous two



**Fig. 7** Illustration of the bundled and ranked list interfaces used in the explorative search study. The interfaces include the search field, a list of previously searched entities, the search results (ranked list or bundles), and a description pane. **a** Bundled search results. **b** Ranked list search results

sections. The interfaces of the two systems are shown in Fig. 7a, b. For each initial entity, we perform a search, and for each result we also perform a search (using result as the query), and similarly for the following results, allowing for a search depth of 3 levels (due to overlap in some result sets, exploration may continue further). To help the understanding of the results, we provide a pane populated with a description of the entity (via Wikipedia) when a cursor is hovered over one, including an image, if available. Both interfaces had the same “look and feel” so that a variability in the user search experience was solely due to whether the results were presented as a ranked list or in bundles.

To gauge the user experience with our systems, we adapt the questions from Arguello et al. (2012) on search effectiveness (*SE*), and O’Brien (2010) concerning user involvement (*IN*), perceived usability (*PU*), and task endurance (*EN*). The questions, in a form of a statement with which the user would rate their agreement on a Likert scale, are listed in Table 11. We use these questions because our focus was not to require answers that could serve as measures of success or task completion, but to assess more subjective aspects of the search experience. In our work, we aimed to evaluate whether users appreciated differently explorative experiences based on different organizations of the results, where user satisfaction is highly subjective, as it depends on the user background, preferences and expertise.

We provided MT workers with the following instructions: “Complete a given search task and answer a survey about it. You are assigned a search task to solve regarding a given topic. After completing your search, you are required to answer some questions about your search experience at the bottom of the page. You are provided with a search engine that

**Table 11** Search experience questionnaire

---

|            |  |
|------------|--|
| <i>EN1</i> | Searching using this system was worthwhile   |
| <i>EN2</i> | I consider my search experience a success  |
| <i>EN3</i> | This search experience did not work out as I had planned   |
| <i>EN4</i> | I would recommend searching using this system to my friends and family                             |
| <i>IN1</i> | I was really drawn into the task   |
| <i>IN2</i> | I felt involved in the task  |
| <i>IN3</i> | The experience was fun   |
| <i>NO1</i> | I continued to use this system out of curiosity  |
| <i>NO2</i> | This system incited my curiosity   |
| <i>NO3</i> | I felt interested in my task   |
| <i>PU1</i> | I felt frustrated while using this system for this task  |
| <i>PU2</i> | I found this system confusing to use   |
| <i>PU3</i> | I felt annoyed while using this system for this task   |
| <i>PU4</i> | I felt discouraged while using this system for this task   |
| <i>PU5</i> | This task was demanding  |
| <i>PU6</i> | I felt in control of my search experience  |
| <i>PU7</i> | I could not do some of the things I needed to do using this system                                 |
| <i>SE1</i> | The system provided enough information to help me solve the search tasks                           |
| <i>SE2</i> | The system provided me with different kinds of information   |
| <i>SE3</i> | The presentation of search results allowed me to easily identify relevant information              |
| <i>SE4</i> | The presentation of search results helped me get an overview of the types of information available |

---

will provide you with results to solve this task. You can query as many things as you want, given that they are included in the list of auto-completions. You will be asked your Worker ID to login to the survey, and will be presented a code upon completion of the survey”. Figure 8 shows the preview of a HIT shown to a worker. We used the same mechanism as before to direct to our servers, and provide her with a code that she could use to claim payment on MT after completing the task.

For each task we aimed to collect 5 annotations. Due to the random assignment and grouping of the tasks on MT, we attained a minimum of 4 annotations, with a mean of 5.13 annotations per task, resulting in 308 tasks completed in total (6, 468 individual questions answered). We were able to recruit 58 workers for this study, yielding a total of 277 distinct workers in all experiments. Users were not allowed to see both interfaces for the same task. Tasks were served to the users by sampling at random from the tasks with missing annotations, excluding those for which the user had already been shown one of the two interfaces and provided the corresponding annotation.

The overlap with the workers involved in the previous experiments was small: only 7 workers (2.5 % of the total) who participated in this last task also participated in the first study, and only 5 (1.8 % of the total) also took part in the second. Only 2 workers (less than 1 % of the total) participated in all the three studies. Due to the highly subjective nature of this task, no gold standard was used. Instead, users whose median time per task was below a minute were removed from the analysis (resulting in 298 tasks remaining for the analysis).

The average scores for the questions are shown in Table 12, along with their  $p$  values. In Table 12, we find an overall more positive response to the bundling interface, with participants agreeing more that using the system was worthwhile (*EN1*), and that they would recommend it to their friends and family (*EN4*). Instead, they felt like the ranking interface was more frustrating (*PU1*), confusing (*PU2*), and discouraging (*PU4*). The opinions on the information presented did not differ (*SE1* and *SE2*), but the presentation of results was favored in the bundling interface, both for identifying relevant information (*SE3*) and providing an overview (*SE4*). Finally, the respondents using bundling interface felt more involved in the task (*IN2*) and felt the task was less demanding (*PU5*).

Novelty (e.g. inciting curiosity) was positive in both systems, with no difference between them. The latter may be caused by the same “look and feel” for both interfaces and the fact that the results shown are comparable, albeit presented in a different way. Both systems incited curiosity, suggesting the suitability of entity search in promoting explorative search. Overall the bundled interface was in general preferred to the ranked list [and

The screenshot shows a web interface titled "HIT Preview". The main content area contains the following text:

**Complete a given search task and answer a survey about it**

You are assigned a search task to solve regarding a given topic. After completing your search, you are required to answer some questions about your search experience at the bottom of the page.

You are provided with a search engine that will provide you with results to solve this task. You can query as many things as you want, given that they are included in the list of auto-completions.

You will be asked your Worker ID to login to the survey, and will be presented a code upon completion of the survey.

Survey link: <http://deesse.illinois-project.eu/study3/index.php>

Provide the survey code here:

**Fig. 8** Instructions shown to the worker

**Table 12** Search experience task mean responses (ranging 1 strongly disagree to 5 strongly agree)

|            | Ranked | Bundles | <i>p</i> |            | Ranked | Bundles | <i>p</i> |
|------------|--------|---------|----------|------------|--------|---------|----------|
| <i>EN1</i> | 2.8    | 3.3     | ***      | <i>PU1</i> | 3.2    | 2.6     | ***      |
| <i>EN2</i> | 3.0    | 3.4     | *        | <i>PU2</i> | 3.1    | 2.3     | ****     |
| <i>EN3</i> | 3.5    | 3.0     | **       | <i>PU3</i> | 3.2    | 2.6     | ***      |
| <i>EN4</i> | 2.4    | 3.2     | ****     | <i>PU4</i> | 3.2    | 2.5     | ****     |
| <i>IN1</i> | 3.6    | 3.8     |          | <i>PU5</i> | 2.5    | 2.0     | **       |
| <i>IN2</i> | 3.8    | 4.1     | *        | <i>PU6</i> | 2.8    | 3.2     | *        |
| <i>IN3</i> | 3.2    | 3.5     | *        | <i>PU7</i> | 3.4    | 2.7     | ****     |
| <i>NO1</i> | 3.4    | 3.4     |          | <i>SE1</i> | 3.1    | 3.3     |          |
| <i>NO2</i> | 3.5    | 3.6     |          | <i>SE2</i> | 4.0    | 4.1     |          |
| <i>NO3</i> | 3.8    | 3.9     |          | <i>SE3</i> | 3.1    | 3.7     | ***      |
|            |        |         |          | <i>SE4</i> | 3.6    | 4.0     | ***      |

*p* < 0.0001\*\*\*\*; *p* < 0.001 \*\*\*;  
*p* < 0.01 \*\*; *p* < 0.05 \*

in the few cases when not, both interfaces were preferred to the same extent), both in terms of search effectiveness and aspects related to engagement. Our results show that users can perform their search tasks effectively, and while doing so, have a positive experience (which for example they want to recommend to others (*EN4*)). This shows the potential of our bundling method in promoting a more engaging explorative search experience.

In Table 13 we further compare the ranked list and the bundling by looking at the correlations between the time and the number of clicks per task, and the scores given to the various questions. The table reports Pearson’s correlations that were statistically significant (*p* < 0.01). The correlations are low, but this is not surprising given the high subjectivity of the tasks, and of users’ behavior. Both with ranked list and bundles, users spent less time and fewer clicks when they had fun, they found relevant information, they felt in control of the experience, and they felt willing to recommend it to others. More demanding and frustrating tasks resulted in more time spent on the task (ranked list) and in more clicks (bundles). If we assume that a better presentation results in less effort (less time and clicks per task) and higher user satisfaction, thus looking at Table 14, which shows that on average, users spent less time and fewer clicks on tasks where results were bundled, we take these results as further testament of the effectiveness of our bundling method.

Also, from Table 14 we can observe that people mostly took around 2 or 3 min to solve the tasks proposed in this study. To gain a better idea whether these times can be considered compatible with explorative activities (independently from whether the user is shown a ranked list of a bundled result set, as we are just looking at the nature of the tasks here), we compared this task duration with average session length in a large fragment of the Yahoo search log, spanning the same time interval as the Yahoo Answers dataset from which the entity network was built. When looking at physical sessions built simply using the traditional rule of 30-min maximum timeout between two consecutive actions to break the activity of users into segments, we observed an average session length of 3 min, somewhat in line with the query log analysis literature, which reports average session durations of 5–12 min (He et al. 2002; Jansen et al. 2007).

However, physical sessions typically contain many different activities from the same user, whereas in our study users are asked to focus on a single and well-defined task. For a more proper comparison, we thus segmented the physical sessions of our query log into *logical* sessions or *missions*, i.e. topically coherent fragments of sessions where users are focused on a single information need. We found that missions are typically very short:

**Table 13** Correlation statistics

| Ranked list |                  |                  |                  |                  |                  |
|-------------|------------------|------------------|------------------|------------------|------------------|
| Time        | <i>PU1</i> 0.22  | <i>PU2</i> 0.20  | <i>PU3</i> 0.19  | <i>PU4</i> 0.20  | <i>PU5</i> 0.37  |
|             | <i>IN3</i> -0.17 | <i>PU6</i> -0.18 | <i>SE3</i> -0.23 |                  |                  |
| Clicks      | <i>PU7</i> -0.17 | <i>SE2</i> 0.17  |                  |                  |                  |
| Bundles     |                  |                  |                  |                  |                  |
| Time        | <i>EN1</i> -0.22 | <i>EN4</i> -0.19 | <i>IN1</i> -0.20 | <i>IN3</i> -0.16 | <i>SE1</i> -0.16 |
| Clicks      | <i>EN1</i> -0.20 | <i>EN2</i> -0.19 | <i>EN4</i> -0.22 | <i>PU6</i> -0.22 | <i>SE1</i> -0.24 |
|             | <i>EN3</i> 0.20  | <i>PU3</i> 0.19  | <i>PU4</i> 0.16  | <i>PU7</i> 0.22  |                  |

**Table 14** Time and clicks per ranked lists and bundled results

| Method      | Time   |     |      | Clicks |     |     |
|-------------|--------|-----|------|--------|-----|-----|
|             | Median | Avg | Max  | Median | Avg | Max |
| ranked list | 154    | 194 | 1179 | 1      | 1   | 11  |
| bundles     | 130    | 166 | 1681 | 0      | 0.6 | 10  |

average mission length in the query log was 12 s; 60 % of the sessions are shorted than 26 s, and 70 % shorter than 74 s. The fact that the times measured in this experiment are much longer than average search times observed in the search engine log suggests that they are more in line with an explorative kind of activity.

## 10 Discussion

We discuss our results as well as putting them in the context of other works and future investigations.

### 10.1 Extending explorative entity search with composite retrieval

This work describes an evaluation of the application of the *composite retrieval* paradigm to the general-purpose *exploratory entity search* task, and as such is a novel contribution to the enriched-search literature. Unlike in previous studies (Stamou and Kozanidis 2009; Yogev et al. 2012), our system allows non-expert, every-day users to browse a social-media collection such as Yahoo Answers not at document, but at *entity* level, aided by high-level topical bundling. Our findings support those on image (Yee et al. 2003), web (Wu et al. 2003), and biomedical text (Pratt and Fagan 2000) search, indicating that topical category hierarchies are beneficial in the consumption of results. In fact, we show that the information itself is not perceived differently, but that the interface (bundles of entities as opposed to a ranked list of entities) provides a better overview of information available, and allows users to more easily locate relevant results.

Our results enrich the existing work on entity search, demonstrating that entity search can go beyond the standard ranked list of results. We have shown this to be the case for the



*explorative* search scenario. It would be interesting to see how composite retrieval benefits other search scenarios.

## 10.2 The potential of social media

Our previous work (Bordino et al. 2013a) showed that users perceive result entities as answers to their queries more positively when the entities and their relationships are extracted from a large uncurated Q and A forum such as Yahoo Answers, rather than a more curated social media such as Wikipedia. In this work, users perceived the bundled entities even better. This opens a new area of research on how to better exploit social media platforms (and not just Yahoo Answers) to promote explorative search, using a more complex presentation to support an engaging interaction.

Our method is completely generalizable to other social media beyond Yahoo Answers. Our system leverages the abstraction of an entity network to build a general and powerful representation of the content produced and consumed by media users. Entities and relations are extracted from the text by using standard natural-language processing and information retrieval techniques, and the additional features used to enrich the network, such as sentiment and topical categories, can also be derived from the text itself (although in the case of Yahoo Answers we exploited the explicit built-in categorization of questions and answers to extract topical features for entities). Thus our whole system can be built from other social media exposing content that is actively produced and shared among users. Our own previous work in fact compared Yahoo Answers with Wikipedia. It would be interesting to extend the analysis to other media as well.

## 10.3 Test the bundles

Although results are encouraging for some of our bundling algorithms, our results show that some can be detrimental. Thus, it is important to not do bundling just for the sake of it. Interestingly we find that the simplest method, i.e., the one using the categories of the documents (questions and answers) associated with the query entity, performed the best. This illustrates a successful use of user-generated content for information structuring, extending earlier work on data-driven organization (Chen and Dumais 2000; Cutting et al. 1992).

## 10.4 Exploring the universe of bundling methods

Topical categories are not the only metadata available for bundling of results. In this work we explored the usefulness of sentiment polarity in result bundling, and found it lacking. However, when combined with topical categories in a certain way, it was still preferred over the baseline (see *categories-then-polarity* results in Table 9). It is possible that sentiment is helpful in specific tasks involving controversial topics. For instance, in the case of *Euthanasia in the Netherlands* we find the positive bundle to include *Dignity in Dying* and *Advance health care*, the negative one *Infanticide* and *Lethal injection*, and the neutral one names of universities. Other document metadata like temporal distribution (for time-sensitive or event queries), text quality (for readability and style), and linking to outside material are all candidates for further development of bundles.

## 10.5 Limitations of crowdsourcing platforms

We evaluate our six bundling algorithms both as static retrieved result sets, as well as in an interactive search. Although crowdsourcing platforms enable the recruitment of hundreds of participants, the interpretation of their feedback can be challenging. In the first two studies, we compute the inter-annotator agreement as percentage overlap between labels, and in some cases we show it to be as low as 65 %. The reason for this amount of variation may be due to different backgrounds of the labelers, interface preferences, or other outside factors. In the first two studies, we obtain low values for the Krippendorff's alpha coefficient. In study 3 we attempt to further understand qualitatively the user experiences during the search tasks. This highly subjective task results in a maximum agreement of 0.22, as computed using inter-annotator agreement designed for crowdsourced tasks (Snow et al. 2008). Further studies employing free-form feedback would shed light on further experiences with these systems.

Recent works (Ribeiro et al. 2011; Keimel et al. 2012; Hanhart et al. 2014) have increasingly used crowdsourcing for complex subjective tasks, like gathering quality feedback for images or videos, showing that despite the well-known limitations, crowdsourcing experiments can deliver accurate and repeatable results, yielding high correlations with subjective evaluations obtained in controlled laboratory environments.

For instance, Hofeld et al. (2013) analyze in depth the usage of crowdsourcing user studies for collecting subjective feedback and quality-of-experience assessments, identifying a set of *best practices* for reliability, suggesting to incorporate reliability controls such as verification tests (like captchas or computation of simple text equations), consistency tests or questions about the test contents, gold-standard data and application-layer monitoring (monitoring of response time of users and browser events to capture the level of focus). The authors recommend to include diverse reliability items but not too many, not to incur in the risk that the assessment becomes lengthy and the users drop the survey. In our work, we have employed several of these recommended practices, selecting workers with good credentials on the platform, using golden-truth questions and monitoring the time spent on the tasks, dropping the contributes of workers that were too quickly in performing the user study (those users whose median time spent on a task was below a minute).

Nonetheless we are aware that a lab study with physical people would allow a much fine-grained control on the quality of the assessments, and we plan to do this in future work.

## 11 Conclusions and future work

This work shows that topical bundling is indeed beneficial to explorative entity search. Our system, built on top of a large Q&A dataset from Yahoo Answers, provides six alternative result-bundling algorithms based on the topical categories of the query entity, the categories shared by the result entities, the sub-topics of the query entity (identified using search-log data), and aggregated sentiment of the documents in which the entity is present. In three crowdsourced studies we show the benefits of topical bundling in the way users perceive and understand the results.

We hope this work encourages further design and evaluation of such systems. Future exploration of metadata associated with the entities extracted from the text in which they

occur, as well as careful pairing with the appropriate search tasks, would allow the use of the underlying dataset beyond the entity similarity computation used for retrieval.

In parallel, behavioral data could be exploited to derive useful features for personalizing results towards users consider interesting based on their profile and activities.

The design of appropriate visualization techniques is necessary to ensure an engaging search experience with bundled results. Further qualitative evaluation efforts are needed to understand the relationship between the bundles and the diversity of search users.

**Acknowledgments** We are very thankful to Byungkyu Kang for his help in designing the explorative-search user study.

**Funding** This work was partially funded by LiMoSiNe project ([www.limosine-project.eu](http://www.limosine-project.eu)).

**Compliance with ethical standards**

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

- Adar, E., Teevan, J., Agichtein, E., & Maarek, Y. (Eds.). (2012). *Proceedings of the fifth international conference on web search and web data mining, WSDM'2012*. Seattle, WA: ACM.
- Amer-Yahia, S., Bonchi, F., Castillo, C., Feuerstein, E., Méndez-Díaz, I., & Zabala, P. (2014). Composite retrieval of diverse and complementary bundles. *IEEE Transactions on Knowledge and Data Engineering*, 26(11), 2662–2675. doi:[10.1109/TKDE.2014.2306678](https://doi.org/10.1109/TKDE.2014.2306678).
- Amitay, E., Carmel, D., Har'El, N., Ofek-Koifman, S., Soffer, A., & Yogev, S. et al. (2009). Social search and discovery using a unified approach. In C. Cattuto, G. Ruffo & F. Menczer (Eds.). *Proceedings of the 20th ACM conference on hypertext and hypermedia (HYPERTEXT'2009)* (pp. 199–208). Torino: ACM. June 29–July 1, 2009. doi:[10.1145/1557914.1557950](https://doi.org/10.1145/1557914.1557950).
- Angel, A., Chaudhuri, S., Das, G., & Koudas, N. (2009). Ranking objects based on relationships and fixed associations. In M. L. Kersten, B. Novikov, J. Teubner, V. Polutin & S. Manegold (Eds.). *Proceedings of ACM 12th international conference on extending database technology (EDBT'2009)*. *ACM international conference proceeding series* (Vol. 360, pp. 910–921). Saint Petersburg, March 24–26, 2009. doi:[10.1145/1516360.1516464](https://doi.org/10.1145/1516360.1516464).
- Arguello, J., Wu, W., Kelly, D., & Edwards, A. (2012). Task complexity, vertical display and user interaction in aggregated search. In W. Hersh et al. (Eds.) (pp. 435–444) 2012. doi:[10.1145/2348283.2348343](https://doi.org/10.1145/2348283.2348343).
- Baeza-Yates, R. A. (2010). Searching the web of objects. In A. Dearle & R. Zicari (Eds.). *Proceedings of objects and databases: Third international conference (ICOODB'2010)*. *Lecture notes in computer science* (Vol. 6348, pp. 6–7). Frankfurt/Main: Springer. September 28–30, 2010. doi:[10.1007/978-3-642-16092-9\\_2](https://doi.org/10.1007/978-3-642-16092-9_2).
- Balog, K., Meij, E. J., & de Rijke, M. (2010). Entity search: Building bridges between two worlds. In *Proceedings of the 3rd international semantic search workshop (SEMSEARCH'10)* (pp. 1–5). New York, NY: ACM.
- Baraglia, R., Morales, G. D. F., & Lucchese, C. (2010). Document similarity self-join with mapreduce. In G. I. Webb et al. (Eds.) (pp. 731–736) 2010. doi:[10.1109/ICDM.2010.70](https://doi.org/10.1109/ICDM.2010.70).
- Boldi, P., Bonchi, F., Castillo, C., Donato, D., Gionis, A., & Vigna, S. (2008). The query-flow graph: Model and applications. In J. G. Shanahan et al. (Eds.) (pp. 609–618) 2008. doi:[10.1145/1458082.1458163](https://doi.org/10.1145/1458082.1458163).
- Boldi, P., Bonchi, F., Castillo, C., & Vigna, S. (2009). From “dango” to “japanese cakes”: Query reformulation models and patterns. In *Main conference proceedings of 2009 IEEE/WIC/ACM international conference on web intelligence (WI'2009)* (pp. 183–190). Milan: IEEE Computer Society. September 15–18, 2009. doi:[10.1109/WI-IAT.2009.34](https://doi.org/10.1109/WI-IAT.2009.34).
- Bonchi, F., Perego, R., Silvestri, F., Vahabi, H., & Venturini, R. (2012). Efficient query recommendations in the long tail via center-piece subgraphs. In W. Hersh et al. (Eds.) (pp. 345–354) 2012. doi:[10.1145/2348283.2348332](https://doi.org/10.1145/2348283.2348332).
- Bordino, I., Mejova, Y., & Lalmas, M. (2013a). Penguins in sweaters, or serendipitous entity search on user-generated content. In Q. He et al. (Eds.) (pp. 109–118) 2013. doi:[10.1145/2505515.2505680](https://doi.org/10.1145/2505515.2505680).

- Bordino, I., Morales, G. D. F., Weber, I., & Bonchi, F. (2013b). From machu\_picchu to “rafting the urubamba river”: Anticipating information needs via the entity-query graph. In S. Leonardi, A. Panconesi, P. Ferragina & A. Gionis (Eds.). *Sixth ACM international conference on web search and data mining (WSDM'2013)* (pp. 275–284). Rome: ACM. doi:10.1145/2433396.2433433.
- Bordino, I., Van Laere, O., Lalmas, M., & Mejova, Y. (2014). Driving curiosity in search with large-scale entity networks. *SIGWEB Newsletter (Autumn)*, doi:10.1145/2682914.2682919.
- Borlund, P., & Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 53(3), 225–250. doi:10.1108/EUM000000007198.
- Bota, H., Zhou, K., & Jose, J. M. (2015). Exploring composite retrieval from the users’ perspective. In A. Hanbury, G. Kazai, A. Rauber & N. Fuhr (Eds.). *Advances in information retrieval: 37th European conference on IR research (ECIR'2015). Proceedings, lecture notes in computer science* (Vol. 9022, pp. 13–24). Vienna. March 29–April 2, 2015. doi:10.1007/978-3-319-16354-3\_2.
- Bota, H., Zhou, K., Jose, J. M., & Lalmas, M. (2014). Composite retrieval of heterogeneous web search. In C. Chung, A. Z. Broder, K. Shim & T. Suel (Eds.). *23rd international world wide web conference (WWW'14)* (pp. 119–130). Seoul: ACM. April 7–11, 2014. doi:10.1145/2566486.2567985.
- Cao, T., Nguyen, Q., Nguyen, A., & Le, T. (2011). Integrating open data and generating travel itinerary in semantic-aware tourist information system. In D. Taniar, E. Pardede, H. Nguyen, J. W. Rahayu & I. Khalil (Eds.). *The 13th international conference on information integration and web-based applications and services (iiWAS'2011)* (pp. 214–221). Ho Chi Minh City: ACM. December 5–7, 2011. doi:10.1145/2095536.2095573.
- Capannini, G., Nardini, F. M., Perego, R., & Silvestri, F. (2011). Efficient diversification of web search results. *PVLDB* 4(7):451–459. <http://www.vldb.org/pvldb/vol4/p451-capannini.pdf>.
- Capra, R., Velasco-Martin, J., & Sams, B. (2011). Collaborative information seeking by the numbers. In *Proceedings of the 3rd international workshop on collaborative information retrieval (CIR'11)* (pp. 7–10). New York, NY: ACM. doi:10.1145/2064075.2064078.
- Chakrabarti, K., Ganti, V., Han, J., & Xin, D. (2006). Ranking objects based on relationships. In S. Chaudhuri, V. Hristidis & N. Polyzotis (Eds.). *Proceedings of the ACM SIGMOD international conference on management of data* (pp. 371–382). Chicago, IL: ACM. June 27–29, 2006. doi:10.1145/1142473.1142516.
- Chen, H., & Dumais, S. T. (2000). Bringing order to the web: Automatically categorizing search results. In T. Turner & G. Szwillus (Eds.). *Proceedings of the CHI 2000 conference on human factors in computing systems* (pp. 145–152). The Hague: ACM. April 1–6, 2000. doi:10.1145/332040.332418.
- Cheng, T., Yan, X., & Chang, K. C. (2007). Entityrank: Searching entities directly and holistically. In C. Koch, J. Gehrke, M. N. Garofalakis, D. Srivastava, K. Aberer & A. Deshpande et al. (Eds.). *Proceedings of the 33rd international conference on very large data bases* (pp. 387–398). University of Vienna, ACM. September 23–27, 2007. <http://www.vldb.org/conf/2007/papers/research/p387-cheng.pdf>.
- Craswell, N., & Szummer, M. (2007). Random walks on the click graph. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr & N. Kando (Eds.). *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'2007)* (pp. 239–246). Amsterdam: ACM. July 23–27, 2007. doi:10.1145/1277741.1277784.
- Cutting, D. R., Pedersen, J. O., Karger, D. R., & Tukey, J. W. (1992). Scatter/gather: A cluster-based approach to browsing large document collections. In N. J. Belkin, P. Ingwersen & A. M. Pejtersen (Eds.). *Proceedings of the 15th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 318–329). Copenhagen: ACM. June 21–24, 1992. doi:10.1145/133160.133214.
- Deng, T., Fan, W., & Geerts, F. (2012). On the complexity of package recommendation problems. In: M. Benedikt, M. Krötzsch, & M. Lenzerini (Eds.). *Proceedings of the 31st ACM SIGMOD–SIGACT–SIGART symposium on principles of database systems (PODS'2012)* (pp. 261–272) Scottsdale, AZ: ACM. May 20–24, 2012. doi:10.1145/2213556.2213592.
- Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4(1), 95–104.
- Evans, B. M., & Chi, E. H. (2008). Towards a model of understanding social search. In B. Begole & D. W. McDonald (Eds.). *Proceedings of the 2008 ACM conference on computer supported cooperative work (CSCW'2008)* (pp. 485–494). San Diego, CA: ACM. November 8–12, 2008. doi:10.1145/1460563.1460641.
- Ferragina, P., & Gulli, A. (2004). The anatomy of snaket: A hierarchical clustering engine for web-page snippets. In J. Boulicaut, F. Esposito, F. Giannotti & D. Pedreschi (Eds.). *Knowledge discovery in databases: 8th European conference on principles and practice of knowledge discovery in databases*,

- (PKDD'2004). *Proceedings of lecture notes in computer science* (Vol. 3202, pp. 506–508). Pisa: Springer. September 20–24, 2004. doi:[10.1007/978-3-540-30116-5\\_48](https://doi.org/10.1007/978-3-540-30116-5_48).
- Fujimura, K., Toda, H., Inoue, T., Hiroshima, N., Kataoka, R., & Sugizaki, M. (2006) Blogranger: A multi-faceted blog search engine. In *Proceedings of the 3rd annual WWE*.
- Gamon, M., Yano, T., Song, X., Apacible, J., & Pantel, P. (2013). Identifying salient entities in web pages. In Q. He et al. (Eds.) (pp. 2375–2380) 2013. doi:[10.1145/2505515.2505602](https://doi.org/10.1145/2505515.2505602).
- Grassi, M., Cambria, E., Hussain, A., & Piazza, F. (2011). Sentic web: A new paradigm for managing social media affective information. *Cognitive Computation*, 3(3), 480–489. doi:[10.1007/s12559-011-9101-8](https://doi.org/10.1007/s12559-011-9101-8).
- Guo, X., Xiao, C., & Ishikawa, Y. (2012). Combination skyline queries. *Trans Large-Scale Data- and Knowledge-Centered Systems*, 6, 1–30. doi:[10.1007/978-3-642-34179-3\\_1](https://doi.org/10.1007/978-3-642-34179-3_1).
- Hanhart, P., Korshunov, P., & Ebrahimi, T. (2014). Crowd-based quality assessment of multiview video plus depth coding. In *2014 IEEE international conference on image processing (ICIP'2014)* (pp. 743–747). Paris: IEEE. October 27–30, 2014. doi:[10.1109/ICIP.2014.7025149](https://doi.org/10.1109/ICIP.2014.7025149).
- He, D., Göker, A., & Harper, D. J. (2002). Combining evidence for automatic web session identification. *Information Processing and Management*, 38(5), 727–742. doi:[10.1016/S0306-4573\(01\)00060-7](https://doi.org/10.1016/S0306-4573(01)00060-7).
- He, Q., Iyengar, A., Nejd, W., Pei, J., & Rastogi, R. (Eds.). (2013). In *22nd ACM international conference on information and knowledge management (CIKM'13)*. San Francisco, CA: ACM. October 27–November 1, 2013. <http://dl.acm.org/citation.cfm?id=2505515>
- Held, C., & Cress, U. (2009). Learning by foraging: The impact of social tags on knowledge acquisition. In U. Cress, V. Dimitrova & M. Specht (Eds.). *Learning in the synergy of multiple disciplines, 4th European conference on technology enhanced learning (EC-TEL'2009). Proceedings of the lecture notes in computer science* (Vol. 5794, pp. 254–266). Nice: Springer. September 29–October 2, 2009. doi:[10.1007/978-3-642-04636-0\\_24](https://doi.org/10.1007/978-3-642-04636-0_24).
- Hersh, W. R., Callan, J., Maarek, Y., & Sanderson, M. (Eds.). (2012). *The 35th international ACM SIGIR conference on research and development in information retrieval (SIGIR'12)*, Portland, OR: ACM. August 12–16, 2012. <http://dl.acm.org/citation.cfm?id=2348283>.
- Hofeld, T., Keimel, C., Hirth, M., Gardlo, B., Habigt, J., & Diepold, K., et al. (2013). Crowdstesting: A novel methodology for subjective user studies and QoE evaluation. Technical report 486, Department of Computer Science.
- Hoffart, J., Yosef, M. A., Bordino, I., Fürstenauf, H., Pinkal, M., & Spaniol, M., et al. (2011). Robust disambiguation of named entities in text. In *Proceedings of the 2011 conference on empirical methods in natural language processing (EMNLP'2011). A meeting of SIGDAT, a special interest group of the ACL* (pp. 782–792). Edinburgh: ACL, John McIntyre Conference Centre. July 27–31, 2011. <http://www.aclweb.org/anthology/D11-1072>.
- Iaquinta, L., de Gemmis, M., Lops, P., Semeraro, G., Filannino, M., & Molino, P. (2008). Introducing serendipity in a content-based recommender system. In F. Xhafa, F. Herrera, A. Abraham, M. Köppen & J. M. Benítez (Eds.). *8th International conference on hybrid intelligent systems (HIS'2008)* (pp. 168–173). Barcelona: IEEE Computer Society. September 10–12, 2008. doi:[10.1109/HIS.2008.25](https://doi.org/10.1109/HIS.2008.25).
- Jansen, B. J., & Pooch, U. W. (2001). A review of web searching studies and a framework for future research. *JASIST*, 52(3), 235–246. doi:[10.1002/1097-4571\(2000\)9999:9999<AID-ASII1607>3.0.CO;2-F](https://doi.org/10.1002/1097-4571(2000)9999:9999<AID-ASII1607>3.0.CO;2-F).
- Jansen, B. J., Spink, A., Blakely, C., & Koshman, S. (2007). Defining a session on web search engines. *JASIST*, 58(6), 862–871. doi:[10.1002/asi.20564](https://doi.org/10.1002/asi.20564).
- Jeh, G., & Widom, J. (2003). Scaling personalized web search. In G. Hencsey, B. White, Y. R. Chen, L. Kovács & S. Lawrence (Eds.). *Proceedings of the twelfth international world wide web conference (WWW'2003)* (pp. 271–279). Budapest: ACM. May 20–24, 2003. doi:[10.1145/775152.775191](https://doi.org/10.1145/775152.775191).
- Käki, M. (2005). Findex: Search result categories help users when document ranking fails. In G. C. van der Veer & C. Gale (Eds.). *Proceedings of the 2005 conference on human factors in computing systems (CHI'2005)* (pp. 131–140). Portland, OR: ACM. April 2–7, 2005. doi:[10.1145/1054972.1054991](https://doi.org/10.1145/1054972.1054991).
- Kashyap, A., & Hristidis, V. (2012). Comprehension-based result snippets. In X. Chen, G. Lebanon, H. Wang & M. J. Zaki (Eds.). *21st ACM international conference on information and knowledge management (CIKM'12)* (pp. 1075–1084). Maui, HI: ACM. October 29–November 02, 2012. doi:[10.1145/2396761.2398405](https://doi.org/10.1145/2396761.2398405).
- Keimel, C., Habigt, J., Horch, C., & Diepold, K. (2012). Qualitycrowd: A framework for crowd-based quality evaluation. In M. Domanski, T. Grajek, D. Karwowski & R. Stasinski (Eds.). *2012 picture coding symposium (PCS'2012)* (pp. 245–248). Krakow: IEEE. May 7–9, 2012. doi:[10.1109/PCS.2012.6213338](https://doi.org/10.1109/PCS.2012.6213338).
- Kucuktunc, O., Cambazoglu, B. B., Weber, I., & Ferhatosmanoglu, H. (2012). A large-scale sentiment analysis for yahoo! Answers. In E. Adar et al. (Eds.) (pp. 633–642) 2012. doi:[10.1145/2124295.2124371](https://doi.org/10.1145/2124295.2124371).

- Kulkarni, S., Singh, A., Ramakrishnan, G., & Chakrabarti, S. (2009). Collective annotation of wikipedia entities in web text. In J. F. Elder, IV, F. Fogelman-Soulié, P. A. Flach & M. J. Zaki (Eds.) *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 457–466). Paris: ACM. June 28–July 1, 2009. doi:[10.1145/1557019.1557073](https://doi.org/10.1145/1557019.1557073).
- Laere, O. V., Bordino, I., Mejova, Y., & Lalmas, M. (2014). DEESSE: entity-driven exploratory and serendipitous search system. In J. Li, X. S. Wang, M. N. Garofalakis, I. Soboroff, T. Suel & M. Wang (Eds.). *Proceedings of the 23rd ACM international conference on conference on information and knowledge management (CIKM'2014)* (pp. 2072–2074). Shanghai: ACM. November 3–7, 2014. doi:[10.1145/2661829.2661853](https://doi.org/10.1145/2661829.2661853).
- Lagun, D., & Agichtein, E. (2011). Viewer: A tool for large-scale remote studies of web search result examination. In D. S. Tan, S. Amershi, B. Begole, W. A. Kelloog & M. Tungare (Eds.). *Proceedings of the international conference on human factors in computing systems (CHI'2011). Extended abstracts volume* (pp. 2035–2040). Vancouver, BC: ACM. May 7–12, 2011. doi:[10.1145/1979742.1979936](https://doi.org/10.1145/1979742.1979936).
- Liu, Y., & Agichtein, E. (2008). On the evolution of the yahoo! answers QA community. In S. Myaeng, D. W. Oard, F. Sebastiani, T. Chua, M. Leong (Eds.). *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'2008)* (pp. 737–738). Singapore: ACM. July 20–24, 2008. doi:[10.1145/1390334.1390478](https://doi.org/10.1145/1390334.1390478).
- Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). Understanding of internal clustering validation measures. In G. I. Webb (Eds.) et al. (pp. 911–916) 2010. doi:[10.1109/ICDM.2010.35](https://doi.org/10.1109/ICDM.2010.35).
- Meij, E., Weerkamp, W., & de Rijke, M. (2012). Adding semantics to microblog posts. In E. Adar et al. (Eds.) (pp. 563–572) 2012. doi:[10.1145/2124295.2124364](https://doi.org/10.1145/2124295.2124364).
- Mihalcea, R., & Csomai, A. (2007). Wikify!: Linking documents to encyclopedic knowledge. In M. J. Silva, A. H. F. Laender, R. A. Baeza-Yates, D. L. McGuinness, B. Olstad & Ø. H. Olsen (Eds.). *Proceedings of the sixteenth ACM conference on information and knowledge management (CIKM'2007)* (pp. 233–242). Lisbon: ACM. November 6–10, 2007. doi:[10.1145/1321440.1321475](https://doi.org/10.1145/1321440.1321475).
- Miliaraki, I., Blanco, R., & Lalmas, M. (2015). From “selena gomez” to “marlon brando”: Understanding explorative entity search. In *Proceedings of the 24th international conference on world wide web, (WWW'2015)* (pp. 765–775) Florence. May 18–22, 2015. doi:[10.1145/2736277.2741284](https://doi.org/10.1145/2736277.2741284).
- Milne, D. N., & Witten, I. H. (2008). Learning to link with wikipedia. In J. G. Shanahan et al. (Eds.) (pp. 509–518) 2008. doi:[10.1145/1458082.1458150](https://doi.org/10.1145/1458082.1458150).
- O'Brien, H. L. (2010). The influence of hedonic and utilitarian motivations on user engagement: The case of online shopping experiences. *Interacting with Computers*, 22(5), 344–352. doi:[10.1016/j.intcom.2010.04.001](https://doi.org/10.1016/j.intcom.2010.04.001).
- Parameswaran, A. G., Venetis, P., & Garcia-Molina, H. (2011). Recommendation systems with complex constraints: A course recommendation perspective. *ACM Transactions on Information Systems*, 29(4), 20. doi:[10.1145/2037661.2037665](https://doi.org/10.1145/2037661.2037665).
- Paranjpe, D. (2009). Learning document aboutness from implicit user feedback and document structure. In D. W. Cheung, I. Song, W. W. Chu, X. Hu & J. J. Lin (Eds.). *Proceedings of the 18th ACM conference on information and knowledge management (CIKM'2009)* (pp. 365–374). Hong Kong: ACM. November 2–6, 2009. doi:[10.1145/1645953.1646002](https://doi.org/10.1145/1645953.1646002).
- Pirolli, P., Schank, P. K., Hearst, M. A., & Diehl, C. (1996). Scatter/gather browsing communicates the topic structure of a very large text collection. In B. A. Nardi, G. C. van der Veer & M. J. Tauber (Eds.) *Proceedings of the conference on human factors in computing systems: Common ground (CHI'96)* (pp. 213–220). Vancouver, BC: ACM. April 13–18, 1996. doi:[10.1145/238386.238489](https://doi.org/10.1145/238386.238489).
- Pratt, W., & Fagan, L. M. (2000). Research paper: The usefulness of dynamically categorizing search results. *JAMIA*, 7(6), 605–617. doi:[10.1136/jamia.2000.0070605](https://doi.org/10.1136/jamia.2000.0070605).
- Ribeiro, F. P., Florêncio, D. A. F., & Nascimento, V. H. (2011). Crowdsourcing subjective image quality evaluation. In B. Macq & P. Schelkens (Eds.). *18th IEEE international conference on image processing (ICIP'2011)* (pp. 3097–3100). Brussels: IEEE. September 11–14, 2011. doi:[10.1109/ICIP.2011.6116320](https://doi.org/10.1109/ICIP.2011.6116320).
- Rose, D. E., & Levinson, D. (2004). Understanding user goals in web search. In S. I. Feldman, M. Uretsky, M. Najork & C. E. Wills (Eds.). *Proceedings of the 13th international conference on world wide web (WWW'2004)* (pp. 13–19). New York, NY: ACM. May 17–20, 2004. doi:[10.1145/988672.988675](https://doi.org/10.1145/988672.988675).
- Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1), 53–65. doi:[10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).
- Roy, S. B., Amer-Yahia, S., Chawla, A., Das, G., & Yu, C. (2010). Constructing and exploring composite items. In A. K. Elmagarmid & D. Agrawal (Eds.). *Proceedings of the ACM SIGMOD international conference on management of data (SIGMOD'2010)* (pp. 843–854). Indianapolis: ACM. June 6–10 2010. doi:[10.1145/1807167.1807258](https://doi.org/10.1145/1807167.1807258).

- Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613–620. doi:10.1145/361219.361220.
- Shanahan, J. G., Amer-Yahia, S., Manolescu, I., Zhang, Y., Evans, D. A., Kolcz, A., et al. (Eds.). (2008). *Proceedings of the 17th ACM conference on information and knowledge management (CIKM'2008)* Napa Valley, CA: ACM. October 26–30, 2008.
- Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast: but is it good? Evaluating non-expert annotations for natural language tasks. In *2008 Proceedings of the conference on empirical methods in natural language processing (EMNLP'2008). A meeting of SIGDAT, a special interest group of the ACL* (pp. 254–263). Honolulu, HI: ACL. October 25–27, 2008. <http://www.aclweb.org/anthology/D08-1027>.
- Spink, A., Jansen, B. J., Wolfram, D., & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. *IEEE Computer*, 35(3), 107–109. doi:10.1109/2.989940.
- Stamou, S., & Kozanidis, L. (2009). Towards faceted search for named entity queries. In L. Chen, C. Liu, X. Zhang, S. Wang, D. Straszunas & S. L. Tomassen et al. (Eds.). *Advances in web and network technologies, and information management, APWeb/WAIM 2009 international workshops: WCMT 2009, RTBI 2009, DBIR-ENQOIR 2009, PAIS 2009, Revised selected papers. Lecture Notes in Computer Science* (Vol. 5731, pp. 100–112). Suzhou: Springer. April 2–4, 2009. doi:10.1007/978-3-642-03996-6\_10.
- Stefanowski, J., & Weiss, D. (2003). Carrot and language properties in web search results clustering. In E. M. Ruiz, J. Segovia & P. S. Szczepaniak (Eds.). *Proceedings of the web intelligence, first international Atlantic web intelligence conference (AWIC'2003). Lecture Notes in Computer Science* (Vol. 2663, pp. 240–249). Madrid: Springer. May 5–6, 2003. doi:10.1007/3-540-44831-4\_25.
- Tan, P., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Boston: Addison-Wesley.
- Toms, E. G. (2000). Serendipitous information retrieval. In *DELOS workshop: Information seeking, searching and querying in digital libraries*. [http://www.ercim.org/publication/ws-proceedings/DelNoe01/3\\_Toms.pdf](http://www.ercim.org/publication/ws-proceedings/DelNoe01/3_Toms.pdf).
- Tong, H., & Faloutsos, C. (2006). Center-piece subgraphs: Problem definition and fast solutions. In T. Eliassi-Rad, L. H. Ungar, M. Craven & D. Gunopulos (Eds.). *Proceedings of the twelfth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 404–413). Philadelphia, PA: ACM. August 20–23, 2006. doi:10.1145/1150402.1150448.
- Tran, Q. T., Chan, C., & Wang, G. (2011). Evaluation of set-based queries with aggregation constraints. In C. Macdonald, I. Ounis & I. Ruthven (Eds.). *Proceedings of the 20th ACM conference on Information and knowledge management (CIKM'2011)* (pp. 1495–1504). Glasgow: ACM. October 24–28, 2011. doi:10.1145/2063576.2063791.
- Walther, M., & Kaiser, M. (2013). Geo-spatial event detection in the twitter stream. In P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. M. Rüger & E. Agichtein et al. (Eds.). *Advances in information retrieval: 35th European conference on IR research (ECIR'2013). Proceedings of lecture notes in computer science* (Vol. 7814, pp. 356–367). Moscow: Springer. March 24–27, 2013. doi:10.1007/978-3-642-36973-5\_30.
- Webb, G. I., Liu, B., Zhang, C., Gunopulos, D., & Wu, X. (Eds.). (2010). In *The 10th IEEE international conference on data mining (ICDM'2010)*. Sydney: IEEE Computer Society. December 14–17, 2010. <http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=5690658>.
- White, R. W., Marchionini, G., & Muresan, G. (2008). Evaluating exploratory search systems: Introduction to special topic issue of information processing and management. *Information Processing and Management*, 44(2), 433–436. doi:10.1016/j.ipm.2007.09.011.
- White, R. W., & Roth, R. A. (2009). *Exploratory search: Beyond the query-response paradigm. Synthesis lectures on information concepts, retrieval, and services*. Morgan & Claypool Publishers. doi:10.2200/S00174ED1V01Y200901ICR003.
- Wildemuth, B. M., & Freund, L. (2012). Assigning search tasks designed to elicit exploratory search behaviors. In *Human-computer information retrieval symposium (HCIR'2012)* (p. 4). Cambridge, MA: ACM. October 4–5, 2012. doi:10.1145/2391224.2391228.
- Wilson, M. L., Kules, B., Schraefel, M. C., & Shneiderman, B. (2010). From keyword search to exploration: Designing future search interfaces for the web. *Foundations and Trends in Web Science*, 2(1), 1–97. doi:10.1561/1800000003.
- Wu, Y. B., Shankar, L., & Chen, X. (2003). Finding more useful information faster from web search results. In *Proceedings of the 2003 ACM CIKM international conference on information and knowledge management* (pp. 568–571). New Orleans, LA: ACM. November 2–8, 2003. doi:10.1145/956863.956975.
- Yee, K., Swearingen, K., Li, K., & Hearst, M. A. (2003). Faceted metadata for image search and browsing. In G. Cockton & P. Korhonen (Eds.). *Proceedings of the 2003 conference on human factors in*

- computing Systems (CHI'2003)* (pp. 401–408). Ft. Lauderdale, FL: ACM. April 5–10, 2003. doi:[10.1145/642611.642681](https://doi.org/10.1145/642611.642681).
- Yogev, S., Roitman, H., Carmel, D., & Zwerdling, N. (2012). Towards expressive exploratory search over entity-relationship data. In A. Mille, F. L. Gandon, J. Misselis, M. Rabinovich & S. Staab (Eds.). *Proceedings of the 21st world wide web conference (WWW'2012)* (pp. 83–92). Lyon: ACM. April 16–20, 2012 (Companion Volume). doi:[10.1145/2187980.2187990](https://doi.org/10.1145/2187980.2187990).
- Yue, Z., Han, S., & He, D. (2012). An investigation of search processes in collaborative exploratory web search. *Proceedings of the American Society for Information Science and Technology*, 49(1), 1–4.
- Zhou, Y., Nie, L., Rouhani-Kalleh, O., Vasile, F., & Gaffney, S. (2010) Resolving surface forms to wikipedia topics. In C. Huang & D. Jurafsky (Eds.). *Proceedings of the 23rd international conference on computational linguistics (COLING'2010)* (pp. 1335–1343). Beijing: Tsinghua University Press. August 23–27, 2010. <http://aclweb.org/anthology/C10-1150>.