Searching for Interestingness in Wikipedia and Yahoo! Answers

Yelena Mejova[1] Ilaria Bordino[2] Mounia Lalmas[3] Aristides Gionis[4*]

[1,2,3]Yahoo! Research Barcelona, Spain [4]Aalto University, Finland
{[1]ymejova, [2]bordino, [3]mounia}@yahoo-inc.com [4]aristides.gionis@aalto.fi

ABSTRACT

In many cases, when browsing the Web, users are searching for specific information. Sometimes, though, users are also looking for something interesting, surprising, or entertaining. *Serendipitous search* puts interestingness on par with relevance. We investigate how interesting are the results one can obtain via serendipitous search, and what makes them so, by comparing entity networks extracted from two prominent social media sites, Wikipedia and Yahoo! Answers.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Serendipity, Exploratory search

1. INTRODUCTION

Serendipitous search occurs when a user with no a priori or totally unrelated intentions interacts with a system and acquires useful information [**?**]. A system supporting such exploratory capabilities must provide results that are *relevant* to the user's current interest, and yet *interesting*, to encourage the user to continue the exploration.

In this work, we describe an entity-driven exploratory and serendipitous search system, based on enriched entity networks that are explored through random-walk computations to retrieve search results for a given query entity. We extract entity networks from two datasets, Wikipedia, a curated, collaborative online encyclopedia, and Yahoo! Answers, a more unconstrained question/answering forum, where the freedom of conversation may present advantages such as opinions, rumors, and social interest and approval.

We compare the networks extracted from the two media by performing user studies in which we juxtapose interestingness of the results retrieved for a query entity, with relevance. We investigate whether interestingness depends on (*i*) the curated/uncurated nature of the dataset, and/or on (*ii*) additional characteristics of the results, such as sentiment, content quality, and popularity.

2. ENTITY NETWORKS

We extract entity networks from (*i*) a dump of the English Wikipedia from December 2011 consisting of 3 795 865 articles, and (*ii*) a sample of the English Yahoo! Answers dataset from 2010/2011, containing 67 336 144 questions and 261 770 047 answers. We use state-of-the-art methods [**?** **?**] to extract entities from the documents in each dataset.

Next we draw an arc between any two entities e_1 and e_2 that co-occur in one or more documents. We assign the arc a weight $w_1(e_1, e_2) = \mathsf{DF}(e_1, e_2)$ equal to the number of such documents (the *document frequency* (DF) of the entity pair).

This weighting scheme tends to favor popular entities. To mitigate this effect, we measure the rarity of any entity e in a dataset by computing its *inverse document frequency* $\mathsf{IDF}(e) = \log(N) - \log(\mathsf{DF}(e))$, where N is the size of the collection, and $\mathsf{DF}(e)$ is the document frequency of entity e. We set a threshold on IDF to drop the arcs that involve the most popular entities. We also rescale the arc weights according to the alternative scheme $w_2(e_1 \to e_2) = \mathsf{DF}(e_1, e_2) \cdot \mathsf{IDF}(e_2)$.

We use Personalized PageRank (PPR) [**?**] to extract the top n entities related to a query entity. We consider two scoring methods. When using the w_2 weighting scheme, we simply use the PPR scores (we dub this method IDF). When using the simpler scheme w_1, we normalize the PPR scores by the global PageRank scores (with no personalization) to penalize popular entities. We dub this method PN.

We enrich our entity networks with metadata regarding sentiment and quality of the documents. Using SentiStrength[1], we extract sentiment scores for each document. We calculate *attitude* and *sentimentality* metrics [**?**] to measure polarity and strength of the sentiment. Regarding quality, for Yahoo! Answers documents we count the number of points assigned by the system to the users, as indication of expertise and thus good quality. For Wikipedia, we count the number of *dispute* messages inserted by editors to require revisions, as indication of bad quality. We derive sentiment and quality scores for any entity by averaging over all the documents in which the entity appears. We use Wikimedia[2] statistics to estimate the popularity of entities.

3. EXPLORATORY SEARCH

We test our system using a set of 37 queries originating from 2010 and 2011 Google Zeitgeist (`www.google.com/zeitgeist`) and having sufficient coverage in both datasets. Using one of the two algorithms – PN or IDF – we retrieve the top five entities from each dataset – YA or WP – for each query. For comparison, we consider setups consisting of 5 random entities. Note that unlike for conventional retrieval, a random baseline is feasible for a browsing task.

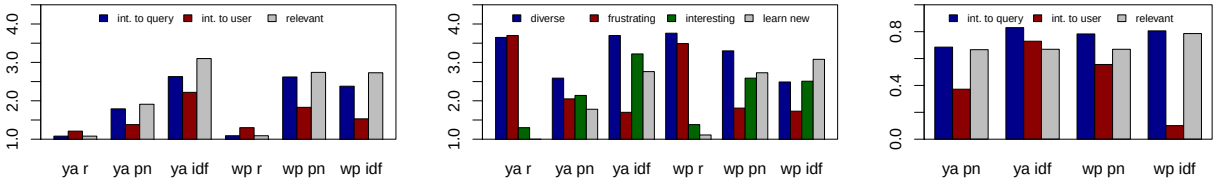
*Work done while the author was at Yahoo! Research.

[1] `sentistrength.wlv.ac.uk`

[2] `dumps.wikimedia.org/other/pagecounts-raw`

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

Figure 1: Performance: (a) and (b) scale range from 1 to 4, (c) correlation range from 0 to 1



(a) Average query/result pair labels (b) Average query set labels (c) Corr. with *learn smth new*

We recruit four editors to annotate the retrieved results, asking them to evaluate each result entity for relevance, interestingness to the query, and interestingness regardless of the query, with responses falling on scale from 1 to 4 (Figure **??**(a)). Both of our retrieval methods outperform the random baseline (at $p < 0.01$). The gain in interestingness to the user despite the query suggests that randomly viewed information is not intrinsically interesting to the user.

Whereas performance improves from PN to IDF for YA, the interestingness to the user is hurt significantly (at $p < 0.01$) for WP (the other measures remain statistically the same). Note that PN uses the weighting scheme w_1, while IDF operates on the networks sparsified and weighted according to function w_2. The frequency-based approach applied by IDF mediates the mentioning of popular entities in a non-curated dataset like YA, but it fails to capture the importance of entities in a domain with restricted authorship.

Next we ask the editors to look at the five results as a whole, measuring *diversity*, *frustration*, *interestingness*, and the ability of the user to *learn something new* about the query. Figure **??**(b) shows that the two random runs are highly diverse but provoke the most frustration. The most diverse and the least frustrating result sets are provided by the YA IDF run. The WP PN run also shows high diversity, but it falls with the IDF constraint. The YA IDF run gives better diversity and interesting scores at $p < 0.01$ than the WP IDF run, while performing statistically the same.

To examine the relationship with the serendipity level of the content, we compute correlation between the *learn something new* label (LSN) and the others. Figure **??**(c) shows the LSN label to be the least correlated with interests of the user in the WP IDF run, and the most for the YA IDF run. Especially in the WP IDF run, the relevance is highly associated with the LSN label. We are witnessing two different searching experiences: in the YA IDF setup the results are diverse and popular, whereas in the WP IDF setup the results are less diverse, and the user may be less interested in the relevant content, but it will be just as educational.

Finally we analyze the metadata collected for the entities in any query-result pair: *Attitude* (A), *Sentimentality* (S), *Quality* (Q), *Popularity* (V), and *Context* (T). For each pair, we calculate the difference between query and result in these dimensions. For Context we compute the cosine similarity between the TF/IDF vectors of the entities. In aggregate, the best connections are between result popularity and relevance (0.234), as well as interestingness of the result to the user (0.227), followed by contextual similarity of result and query (0.214), and quality of the result entity (0.201). These features point to important aspects of a retrieval strategy which would lead to a successful serendipitous search.

Table 1: Retrieval result examples

YA query: **Kim Kardashian**	Attitude	Sentiment.	Quality	Pageviews
Perry Williams	0	0	0	85
Eva Longoria Parker	−0.602	2.018	6	1 450 814
WP query: H1N1 pandemic	**Attitude**	**Sentiment.**	**Quality**	**Pageviews**
Phaungbyin	2	2	1	706
2009 US flu pandemic	1	1	1	21 981

4. DISCUSSION & CONCLUSION

Beyond the aggregate measures of the previous section, the peculiarities of Yahoo! Answers and Wikipedia as social media present unique advantages and challenges for serendipitous search. For example, Table **??** shows potential search YA results for an American socialite *Kim Kardashian*: an actress *Eva Longoria Parker* (whose Wikipedia page has over a million visits in two years), and a footballer *Perry Williams* (who played his last game in 1993). Note the difference in attitude and sentimentality. Yahoo! Answers provides a wider spread of emotion. This data may be of use when searching for potentially serendipitous entities.

Table **??** also shows potential WP results for the query *H1N1 Pandemic*: a town in Burma called *Phaungbyin*, and *2009 flu pandemic in the United States*. We may expect pandemic to be associated with negative sentiment, but the documents in Wikipedia do not display it.

It is our intuition that the two datasets provide a complementary view of the entities and their relations, and that a hybrid system exploiting both resources would provide the best user experience. We leave this for future work.

5. ACKNOWLEDGEMENTS

This work was partially funded by the European Union Linguistically Motivated Semantic Aggregation Engines (LiMoSINe) project[3].

[3] www.limosine-project.eu