Dear Search Engine: What's your opinion about...?

Sentiment Analysis for Semantic Enrichment of Web Search Results

Gianluca Demartini
L3S Research Center
Appelstrasse 9a
30167 Hannover, Germany
demartini@L3S.de

Stefan Siersdorfer
L3S Research Center
Appelstrasse 9a
30167 Hannover, Germany
siersdorfer@L3S.de

ABSTRACT

Search Engines have become the main entry point to Web content, and a large part of the "visible" Web consists in what is presented by them as top retrieved results. Therefore, it would be desirable if the first few results were a representative sample of the entire result set. This paper provides a preliminary study about opinions contained in search engine results for controversial queries such as "cloning" or "immigration". To this end, we extract sentiment metadata from web pages, and compare search engine results for several queries. Furthermore, we compare opinions expressed in the top results to those in other retrieved results to examine whether the top-ranked pages are a good sample of all results from an opinion perspective. In a preliminary empirical analysis, we compare up to 50 results from 3 commercial search engines on 14 controversial queries to study the relation between sentiments, topics, and rankings.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Storage and Retrieval

General Terms

Algorithms, Measurement, Experimentation

Keywords

opinion mining, search result diversification, semantic enrichment

1. INTRODUCTION

Every day news about controversial topics are published on the web. Moreover, people discuss their ideas and opinions in blogs and social web sites. As the amount of Web content is rapidly growing, search engines have become an essential tool for users to find information. For the same reason, the number of pages which are relevant to a query is growing, forcing users to trust the search engine in what it presents them. Moreover, algorithmic results are difficult to control as hundreds of features are involved in the creation of the final ranking. Thus, it may happen that rankings based on popularity of web pages (e.g., PageRank), on topical relevance, and even topic diversity are biased towards a certain opinion.

WWW2010, April 26-30, 2010, Raleigh, North Carolina.
.

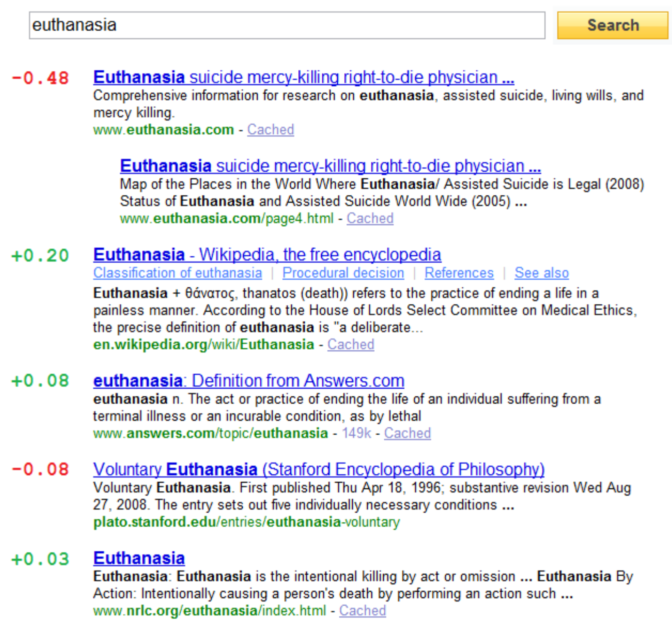


Figure 1: Mock-up search interface showing opinions expressed in top-ranked results.

On the other hand, advances towards richer user interfaces for result presentation have been achieved, and complementary information is added to search interfaces by exploiting available metadata on result pages [9]. In this paper we study an additional source of metadata for web pages by extracting the sentiment in query results. Such additional semantic information can be, for example, displayed to the user within a richer search interface (see Figure 1 for a mock-up visualization interface). Moreover, we can study current search engines from a new angle with the long term goal of producing search results which are more diverse from an opinion perspective.

As web search engine users look mainly at a few top results, it is important to provide a good overview of the entire result set both from a topical point of view (see, for example, [1] on topical diversification of search results) as well as from an opinion point of view which is the focus of this paper. If we want to study diversity of search results we need to understand what the ideal result we expect from a search engine should be. In the case of topic diversification we need to define a measure of diversity (e.g., semantic distance between web pages [5]) and an objective function that, for example, tries to maximize the diversity among pages in the results set. On the other hand, it is not clear what a "good" overview should be if we want to diversify *opinions* in search results. We see four possible "ideal" results that we could compute for a query q:

Balanced Overview. If we assume that the user submits an informational query, then it may be better to show her an objective overview on the topic. A possible option is to present a mixed set of results having both positive (close to +1) and negative (close to -1) documents.

Neutral Overview. With the goal of presenting an objective overview on the topic, the system could also present a result set containing only very objective documents (with sentiment score close to 0, that is, neither positive nor negative).

Realistic Overview. In this case we want to provide the user with a result set which contains an average opinion close to the public one. That is, if we can estimate (e.g., from opinion polls) that 80% of people has a negative opinion about the topic expressed by q, then we might want to show to our user a set of results which contains, for instance in the top five results, 1 positive document and 4 negative ones.

Personalized Overview. If the search engine has information about the user profile it can try to personalize the search results. Given some information about the user opinion, it is possible to select pages for the result set so that her opinion is well reflected.

It can then be a user choice which type of overview he wants to see. As a first step toward opinion diversification in search results, in this paper we focus on:

- Proposing different approaches for computing sentiment metadata associated to web pages.
- Comparing three commercial search engines on the average opinion of the search results they return.
- Comparing the average opinion in top results to those lower in the ranking.

The rest of the paper is structured as follows. In Section 2 we describe previous work on bias in search engines, diversification of search results, and opinion mining. In Section 3 we present methods for extracting sentiment information from web pages both using text classifiers trained on movie reviews as well as a sentiment thesaurus, and aggregate the resulting values to obtain the overall sentiment scores for web pages. Section 4 describes the datasets we used and an experimental comparison of different approaches, search engines, queries, and top-ranked vs. other retrieved results. Section 5 summarizes the conclusions of the paper. In Section 6 we describe possible applications enabled by the ability of extracting opinion metadata from search results.

2. RELATED WORK

An interesting research trend studies bias in search engine results. Proposed techniques compare URLs and content of retrieved results to check whether some *topic* is preferred [10, 11]. That is, the analyzed search engine may show a bias towards a certain type of content or certain URLs. In this paper we aim at performing a similar study which, however, focuses on the *opinions* contained in retrieved results. We want to examine if search engines provide the user only with certain types of opinions by ranking results.

In the context of search result diversification, several approaches deal with the trade-off between having the most relevant and the most diverse result set; see [3] for a survey. In order to provide a diverse set of results there is the need to define a measure of diversity. To this end, measures based on semantic and categorical distance [5, 1] or on novel information [15] have been proposed. Compared to previous work, in this paper we propose ways of measuring *opinions* in web pages which can be used as a different measure of diversity. Our sentiment annotations can be used in combination with standard diversification objectives (see [1, 5, 15, 14]) and existing approximation algorithms for diversifying search results.

Opinion mining approaches have mainly been tested on blog postings where people express their opinion about movies, products, etc. The task of retrieving opinionated posts and their polarity was mainly evaluated in the context of the Text REtrieval Conference[1] (TREC) Blog Track [12]. In order to detect opinionated documents machine learning techniques (e.g., Support Vector Machines [6]) or lexicon-based approaches (based, for instance, on SentiWordNet and the Amazon review data [8]) have been applied. Compared to previous work, we aim at identifying opinions in general web pages which are retrieved for a query by search engines. While using similar techniques, we aggregate sentiment scores in order to study opinion diversity in web search results.

Extracted opinion values from web page content can be seen as additional metadata. In the past years work on enriching search result pages with semantic information such as locations, person names, or dates have been proposed [9]. In this paper we propose methods for computing additional sentiment metadata that can be displayed to the user (see Figure 1 for an example visualization).

3. EXTRACTING SENTIMENT METADATA FROM WEB PAGES

In this section we define possible techniques for mining opinions in web pages which are relevant to a query. We present both approaches relying on text classification using Support Vector Machines (SVMs) as well as based on sentiment thesauri such as the SentiWordNet (SWN) lexicon. Then, we show possible ways of associating a sentiment score in the range [-1,1] to a target web page.

SentiWordNet is a sentimental lexicon built on top of WordNet [4], a thesaurus containing textual descriptions of terms and relationships between them. In SentiWordNet a triple of three *senti values* (pos, neg, obj) (corresponding to positive, negative, or neutral sentiment of a word) are assigned to each set of synonymous words w in WordNet.

[1] `http://trec.nist.gov/`

Such senti values, that were partly assigned by human assessors and partly automatically generated, are in the range of $[0, 1]$, and sum up to 1 for each triple. For instance $(pos, neg, obj) = (0.875, 0.0, 0.125)$ for the term "good" or $(0.25, 0.375, 0.375)$ for the term "ill".

Linear support vector machines construct a hyperplane $\vec{w} \cdot \vec{x} + b = 0$ that separates a set of positive examples (corresponding to positive opinions in our case) from a set of negative examples (negative opinions) with maximum margin. For a new previously unseen document $\vec{d}$, the SVM needs to test whether it lies on the "positive" side or the "negative" side of the space separated by the hyperplane. The decision simply requires computing a scalar product of the vectors $\vec{w}$ and $\vec{d}$. In this paper, we use a standard Bag-of-Words representation of documents with TF weighting and an SVM classifier trained on a polarity dataset [13] consisting of manually labeled movie reviews.

3.1 Estimating sentiment expressed in text

Given a query q and ranked list of n retrieved pages $D = \{d_1, \dots, d_n\}$ we define the sentiment of a page $s(d_i) \in [-1, 1]$ as the opinion expressed in the document about the topic described by q where $s(d_i) = -1$ express a totally negative sentiment and $s(d_i) = +1$ express a totally positive sentiment towards q.

For computing s(d) given a URL returned by a search engine as answer to a query, we first need to get its content. After fetching the HTML page, we adopt state-of-the-art techniques to remove the page template [7] in order to focus on the real content of the document. We can then perform a sentiment analysis on the cleaned-up text.

Lexicon based sentiment extraction. One possible approach is to use a lexical resource such as a set of opinionated words. We can look up each word of the page content in such lexicon for estimating the average opinion of the document (which is assumed to be relevant to the topic). We define $Sent(d)$ as the set of sentences contained in d which can be obtained by standard NLP tools [2]. Then, we compute the average positivity (pos), negativity (neg), and objectivity (obj) of a sentence $st \in Sent(d)$ as:

$$pos(st) = \frac{\sum_{a \in Adjectives(st)} pos(a)}{|Adjectives(st)|} \tag{1}$$

$$neg(st) = \frac{\sum_{a \in Adjectives(st)} neg(a)}{|Adjectives(st)|} \tag{2}$$

$$obj(st) = \frac{\sum_{a \in Adjectives(st)} obj(a)}{|Adjectives(st)|} \tag{3}$$

We take only adjectives into account as they are the best carrier of sentiments.

At this point we can compute the positivity, negativity, and objectivity of a document by computing either the average over all the sentences in the document (SWN_TXT) or just over sentences containing the query terms (SWN_SEN).

Text classification based approach. A second approach is to use binary text classification in order to assign pages to either the positive or the negative class. We build term vectors out of the page d using TF weights and classify it using SVM trained using a set of 10,662 movie reviews manually labeled as positive or negative. Evaluation of the trained classifier by using 80% of the data for training and 20% for testing results in a precision of 0.88 and a recall of 0.81.

We then use the classification results (i.e., the distance of the vector from the hyperplane) $class(d)$ as an estimate of how strong the sentiment is. We can either classify the entire text of the page d (SVM_TXT) or only a document consisting of sentences containing the query terms (SVM_SEN).

3.2 Assigning a sentiment score to a web search result

In the following we list possible simple approaches that we study in this paper to combine sentiment scores in a single semantic annotation for a web page d. We consider:

SWN_SEN_BIN This method estimates $s(d) = +1$ if the average $pos(st)$ is bigger then the average $neg(st)$ for sentences st containing the query terms, $s(d) = -1$ otherwise.

SWN_SEN_MAX This method estimates $s(d) = +1$, $s(d) = 0$, or $s(d) = +1$ if the average $pos(st)$, $obj(st)$, or $neg(st)$ respectively, is maximum when computed on sentences st containing the query terms.

SWN_TXT_BIN This method estimates $s(d) = +1$ if the average $pos(st)$ is bigger then the average $neg(st)$ for all sentences st in d, $s(d) = -1$ otherwise.

SWN_TXT_MAX This method estimates $s(d) = +1$, $s(d) = 0$, or $s(d) = +1$ if the average $pos(st)$, $obj(st)$, or $neg(st)$ respectively, is maximum when computed on all sentences st in d.

SVM_SEN_BIN This method estimates $s(d) = +1$ if the classification result $class(d)$ is bigger than 0, that is, d is classified as positive using only sentences st containing the query terms, $s(d) = -1$ otherwise.

SVM_SEN_REA This method estimates the sentiment of d by normalizing the classification results $class(st)$ in the range [-1,1], $s(d) = norm(class(st))$. The classification is performed only on sentences st containing the query terms.

SVM_TXT_BIN This method estimates $s(d) = +1$ if the classification result $class(d)$ is bigger than 0, that is, d is classified as positive using all sentences st in the document, $s(d) = -1$ otherwise.

SVM_TXT_REA This method estimates the sentiment of d by normalizing the classification results $class(st)$ in the range [-1,1], $s(d) = norm(class(st))$. The classification is performed on the entire web page d.

4. EXPERIMENTAL STUDY

In this section we first describe the dataset we use in this paper. Then, we experimentally compare the proposed algorithms, the different search engines, and queries. Furthermore, we compare top-ranked results to those ranked lower.

Table 1: List of 14 keyword queries considered in this study.

abortion	immigration
anorexia	islam
cloning	marijuana
economy	marriage
employment	nazism
euthanasia	vegetarianism
homosexuality	vividown

4.1 Dataset

We used the search API from three commercial search engines (Google[2], Yahoo![3], and Bing[4]) in order to gather the top 50 retrieved results for 14 controversial queries (see Table 1).

As a first illustrative example we consider the query 'euthanasia' shown in Figure 1. We can see that the first result has a rather negative opinion (scores were obtained using the SVM_TXT_REA algorithm). On that page we can find sentences strongly against 'euthanasia' like "We are committed to the fundamental belief that the intentional killing of another person is wrong." or "There is no quality of life when the patient is dead.".

4.2 Differences between Sentiment Extraction Algorithms

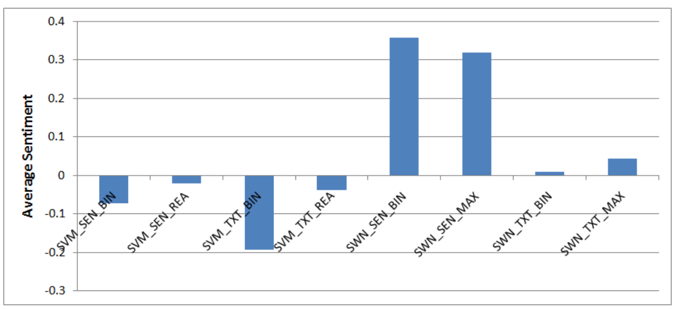


Figure 2: Average sentiment over 14 queries and 3 search engine for different algorithms.

First, we want to compare the output of the different proposed algorithms. Figure 2 shows the average sentiment of our dataset when computed with different algorithms. We see that for both SWN and SVM the sentiment is more positive if we consider only sentences containing the query terms (a t-test between SWN_TXT_BIN and SWN_SEN_BIN returns $p = 0.006$). Moreover, there is a significant difference between different algorithms (ANOVA: $F = 3.41$, $p = 0.003$). This may be explained by the fact that documents try to motivate the positive aspects of a topic in some sentences, while using negative words in other sentences to criticize the opposite. For example, "Vegetarianism is healthy." and "Killing animals is wrong.".

In the following we will present results considering the average sentiment computed using all the different algorithms presented.

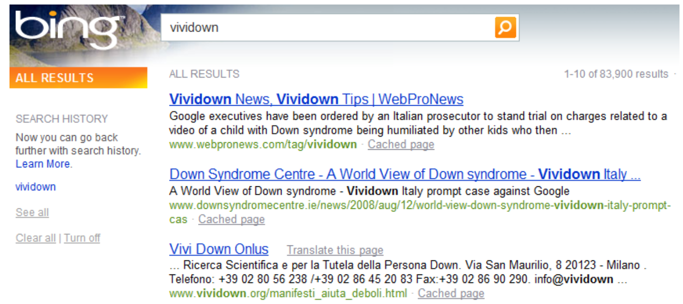
[2] http://www.google.com

[3] http://www.yahoo.com

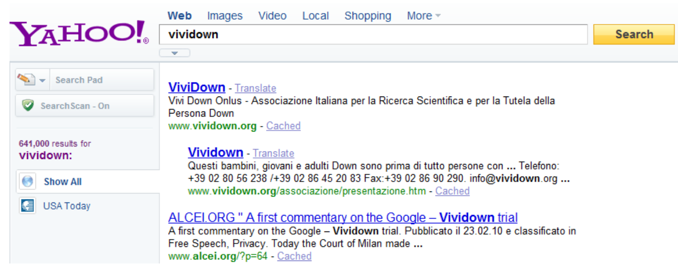
[4] http://www.bing.com

4.3 Differences between Search Engines

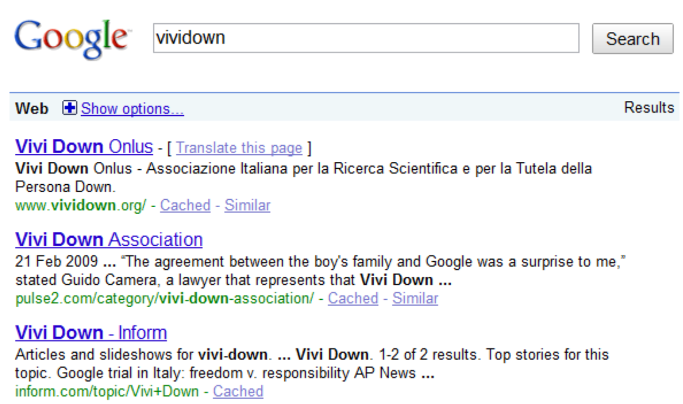
Using the extracted sentiment information, we compared different search engines, and studied possible opinion biases.



(a) Bing.



(b) Yahoo!.



(c) Google.

Figure 3: Top 3 results for the query 'vividown'.

One specific example is related to the query 'vividown'. Recently, Google was convicted in Italy after the charity association ViviDown sued it because of a video being shown on the Google website[5]. We wanted to check whether results returned by Google provide different opinions from those returned by other search engines. In Figure 3 we can see the top 3 results for the query 'vividown' for the three search engines. We can observe that the first result is either the organization home page or an objective page about the trial. The third result in Yahoo!, which is from a website member of the Global Internet Liberty Campaign, talks about the freedom of publishing content on-line and the negative

[5] http://www.reuters.com/article/idUS261896084220100224

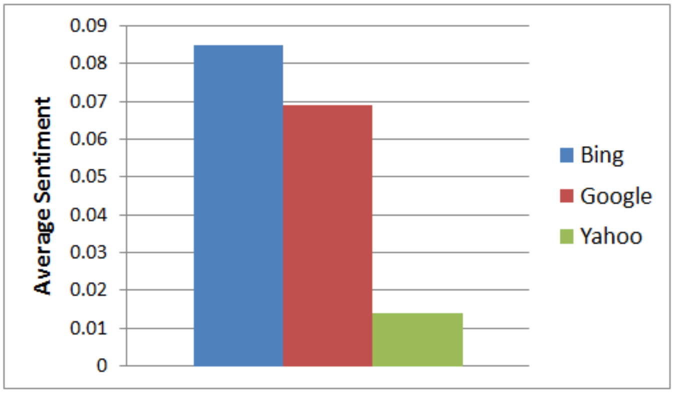


Figure 4: Average Sentiment score in top 5 results for 3 search engines.

repercussions of such a trial using sentences like "the telco market will suffer an alteration of the competition among the various players" and "international players might find Italy a lesser attractive place to do business in". The second Google result reports that "Google officials have not only expressed their solidarity over what happened but have also taken concrete actions". In results 8 and 9 from Google there is, after reporting shortly about the trial, much discussion about the freedom of publishing content on-line and about a different trial where "the Authority accepted that Google was not responsible". On the other hand, top 3 results from Bing are about the ViviDown organization or informative about the trial while the fifth blames the excess of freedom on the current Web. We can conclude that both Google and Yahoo! present results that somehow show the negative aspects of this conviction.

We systematically compared the search engines on the entire set of 14 queries. Figure 4 shows the average sentiment of top 5 results for different search engines. We can see that, in general, scores are very close to 0 with standard deviation values close to 0.1 for *_REA approaches meaning that no extreme sentiments are shown in the top results. Yahoo! presents slightly more objective pages than the other two search engines. Anyway, we could not observe a significant difference between Bing and Yahoo! when looking at top 5 results (t-test: $p = 0.07$). Furthermore, we could not find a significant difference between search engines when comparing the average sentiment of top 50 results over the 14 queries (ANOVA: $F = 0.12$, $p = 0.88$).

We also examined how the average sentiment changes as we go down through the ranked list of results. Figure 5 shows the average sentiment in the top N results for different values of N and different search engines. We can see that, on average, the first result is always more positive than the others. This might be explained by the fact that the fist result is usually the "home page" of the topic which tries to motivate why it is beneficial. On average, the average sentiment at top 1 (i.e., 0.20) is significantly higher (t-test: $p = 0.03$) then that at top 50 (0.05).

4.4 Sentiment for Different Topics

In addition, we compared the overall sentiment expressed by search results for different queries, and, thus, different topics. Figure 6 shows the average sentiment scores for the 14 queries we consider. When we compared the average sentiment of top 50 results for different queries we discovered a significant difference (ANOVA: $F = 20.2$, $p = 7.46E - 11$) showing how results for different queries can contain different sentiments. For instance, comparing a positive query like 'employment' to a query returning negative results like 'marijuana' (see Figure 6) we observed a significant difference (t-test: $p = 0.01$).

Finally, we compared different search engines on specific queries. Figure 7 shows the sentiment of top 5 results for different queries and different search engines. We can see, for example, that while Bing shows a slightly positive opinion about 'vegetarianism', Yahoo! has a slightly negative opinion about it. Other differences can be seen on the negative opinion of Yahoo! about 'marijuana' or on the positive opinion of Google about 'immigration'.

5. CONCLUSION

In this paper we presented an approach for extracting metadata about the sentiments expressed in web search results. To this end, we proposed techniques using both text classification and a lexicon of opinionated words together with different aggregation techniques. Do sentiment and opinions in query results vary for different search engines? What is the relation between specific queries and sentiments? Does sentiment expressed in top retrieved results reflect that in the long tail? These are some of the research questions that we are targeting to gain a better understanding of opinions in search results. Using the extracted sentiment information, we performed a preliminary experimental analysis aiming at comparing opinions expressed for different search engines, and queries in highly ranked and other results. While we observed no significant difference between the three considered search engines, we did observe a difference in the sentiments expressed in retrieved pages for different queries. Moreover, we found a relation between rank and sentiment; for instance, results ranked first by search engines contain, on average, a more positive opinion about the topic expressed by the query than the other results.

6. FUTURE WORK

In this section we describe possible applications enabled by the ability of extracting opinion metadata from search results which will be the focus of our future work.

We aim at developing sentiment diversification measures which could be used, for example, to evaluate effectiveness of current search engines. Moreover, questions about how topic and sentiment diversity interact and whether the sentiment should be seen as a single dimension for diversification, will be studied.

Additional experiments will be performed in order to compare and evaluate against manual judgments different sentiment classification algorithms. In this way we will be able to pick the most appropriate approach for our needs. Moreover, we want to check by means of a user study whether sentiment diversification increases usefulness of search results.

Applications of proposed techniques, which are described in the following, are possible additional directions for our future work.

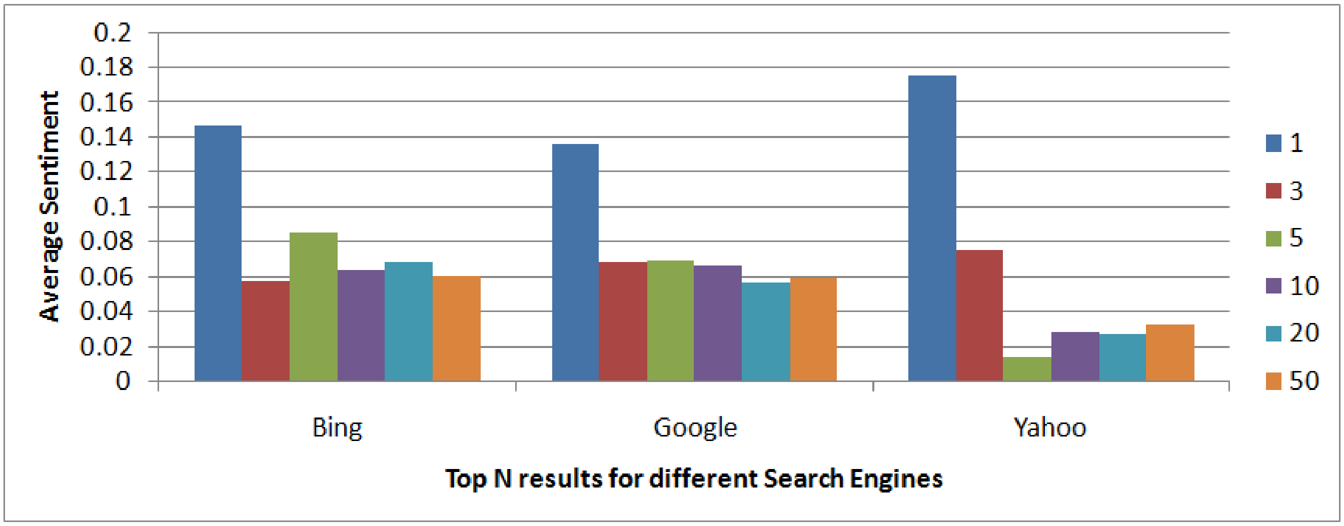


Figure 5: Average Sentiment score in top N results for 3 search engines.

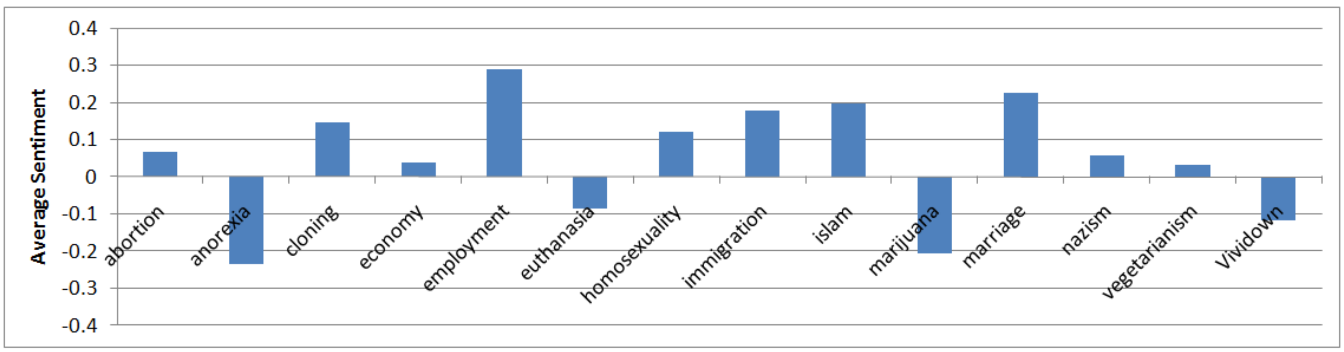


Figure 6: Average Sentiment of top 50 results on the 14 topics considered.

Re-ranking for opinion diversification. After having performed a study of opinions in top results retrieved by search engines, a natural next step is to design techniques that produce a "better" result ranking from the opinion point of view. As described in Section 1 different objectives may be targeted. We aim to develop re-ranking techniques in order to place results on top that provide a good overview of the opinions on the topic. This would mean, for example, to move a result from rank 21 to rank 3 if it contains a different opinion than those expressed in the first 2 results. It is easy to see that this can produce a less relevant result set bringing us to the well known trade-off between relevance and diversity: we want to produce a result set where relevance as well as opinion diversity is maximized. As shown in previous work diversification is an NP-hard problem [1] but approximation functions can be used to produce a result set at query time.

Objectivity of Wikipedia. Given the possibility of measuring the average opinion expressed in web pages, another interesting experiment would be to test the objectivity of Wikipedia (whose pages are often retrieved by search engines in the top results). As it is an encyclopedia, it should contain an objective overview of different topics (included the controversial ones) but, given that content can be created and changed by anyone, this might not be the case for each topic. More then just checking the objectivity of Wikipedia, it might be possible to develop tools that find pages which need revision in order to make them more objective on the described topic.

Sentiment bias of different ranking features. As pointed out in Section 1 many features are currently used by search engines to produce the final ranking of pages. Among them, there are those based on popularity of web pages, document-query similarity measures, or temporal recency of pages. It would be interesting to compare such features in order to understand how they affect opinion bias in search results. Empirical studies could be conducted on large standard collection such as the TREC ClueWeb corpus.

Comparison between search results and public opinion. Do top search engine results reflect public opinion? In order to answer this question we would need to compare the opinion expressed by search results to those of people on the topic. In order to provide an estimate of public opinion on a given topic techniques for mining opinions could be applied to the blogosphere.

Moreover, opinions from different cultural backgrounds could be studied exploiting techniques for automatic trans-

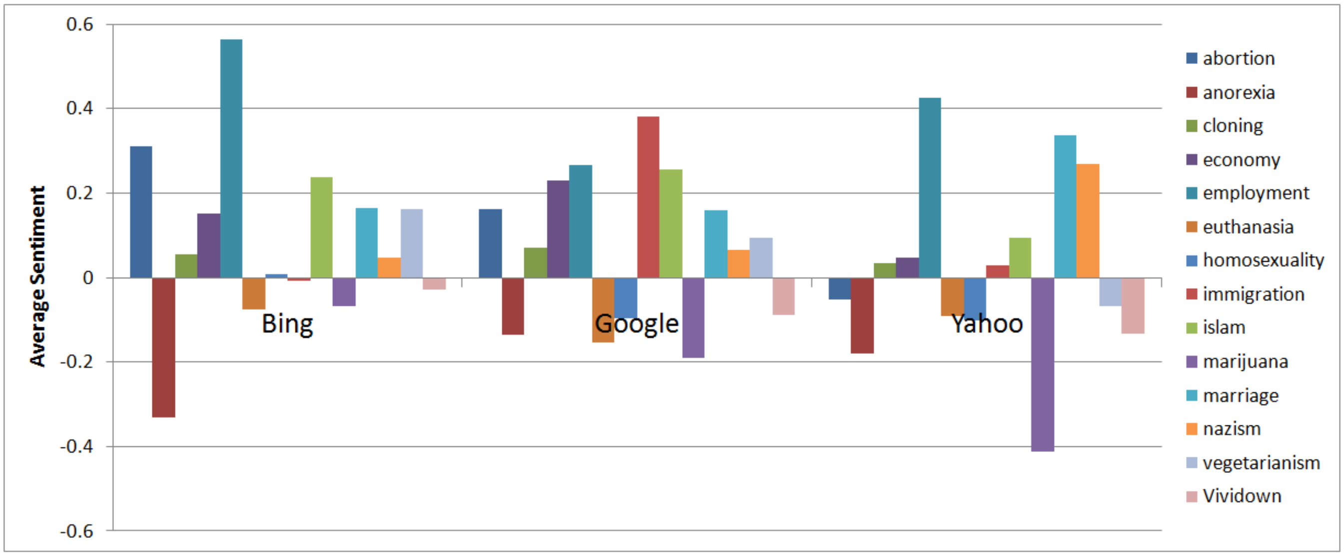


Figure 7: Average sentiment for 14 queries on top 5 results for 3 search engines.

lation of text. This is realistic as existing translation techniques, while not perfect for reflecting complex grammatical constructs, produce good results at a single word level.

Forensic web search. Being able to compute sentiment scores for web pages would enable search for "extreme" documents where the opinion about certain topics is highly positive or negative. This could help governmental institutions to find terror organizations using the web as communication tool. Search engine companies could do a similar job by mining query logs trying to find "extreme" queries, that is, those that generate a highly positive/negative result set.

Acknowledgments.

This work is partially supported by the EU Large-scale Integrating Project LivingKnowledge[6] - Facts, Opinions and Bias in Time (contract no. 231126).

7. REFERENCES

[1] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *WSDM '09*, pages 5–14, New York, NY, USA, 2009. ACM.

[2] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *ACL*, pages 168–175, 2002.

[3] Gianluca Demartini Enrico Minack and Wolfgang Nejdl. Current Approaches to Search Result Diversification. In *Proceedings of First International Workshop on Living Web, Collocated with the 8th International Semantic Web Conference (ISWC-2009), Washington D.C., USA*. CEUR-WS, October 2009.

[4] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[6] http://livingknowledge-project.eu/

[5] Sreenivas Gollapudi and Aneesh Sharma. An axiomatic approach for result diversification. In *WWW '09*, pages 381–390, New York, NY, USA, 2009. ACM.

[6] Lifeng Jia, Clement T. Yu, and Wei Zhang. UIC at TREC 2008 Blog Track. In *TREC*, 2008.

[7] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. Boilerplate detection using shallow text features. In *WSDM '10*, pages 441–450, New York, NY, USA, 2010. ACM.

[8] Yeha Lee, Seung-Hoon Na, Jungi Kim, Sang-Hyob Nam, Hun-Young Jung, and Jong-Hyeok Lee. KLE at TREC 2008 Blog Track: Blog Post and Feed Retrieval. In *TREC*, 2008.

[9] P. Mika. Microsearch: An Interface for Semantic Search. In *ESWC 2008 Workshop: Semantic Search (SemSearch2008)*, Tenerife, Spain, 2008.

[10] Abbe Mowshowitz and Akira Kawaguchi. Assessing bias in search engines. *Inf. Process. Manage.*, 38(1):141–156, 2002.

[11] Abbe Mowshowitz and Akira Kawaguchi. Measuring search engine bias. *Inf. Process. Manage.*, 41(5):1193–1205, 2005.

[12] I. Ounis, C. Macdonald, and I. Soboroff. Overview of the TREC 2008 Blog Track. In *TREC*, 2008.

[13] Bo Pang and Lillian Lee. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL '05*, pages 115–124, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[14] Erik Vee, Utkarsh Srivastava, Jayavel Shanmugasundaram, Prashant Bhat, and Sihem Amer Yahia. Efficient computation of diverse query results. In *ICDE '08*, pages 228–236, Washington, DC, USA, 2008. IEEE Computer Society.

[15] Cheng Xiang Zhai, William W. Cohen, and John Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR '03*, pages 10–17, New York, NY, USA, 2003. ACM.