Recall Oriented Search on the Web using Semantic Annotations

Rianne Kaptein
TNO, The Netherlands
rianne.kaptein@tno.nl

Egon L. van den Broek
TNO, The Netherlands
vandenbroek@acm.org

Gijs Koot
TNO, The Netherlands
gijs.koot@tno.nl

Mirjam A. A. Huis in 't Veld
TNO, The Netherlands
mirjam.huisintveld@tno.nl

ABSTRACT

Web search engines are optimized for early precision, which makes it difficult to perform recall oriented tasks with them. In this article, we propose several ways to leverage semantic annotations and, thereby, increase the efficiency of recall oriented search tasks, with a focus on forensic investigation. Semantic annotations, such as temporal annotations, named entities, and domain context, can be used to rerank, and cluster search result sets. In addition, domain context can be used to improve recall.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information search and retrieval

Keywords

Recall Oriented Search; Named Entity Recognition; Clustering; Domain Context

1. INTRODUCTION

Most searches performed on the Web either target home pages or are informational tasks, which both can be fulfilled with a limited amount of search results. Consequently, web search engines are optimized to excel on these tasks. Their performance is measured in terms of early precision and Normalized Discounted Cumulated Gains (NDCG), focusing on the quality of the highly ranked search results. Their interface as well as their search results are optimized for this; only the first 10 or 20 search results are shown on the search results page, and great care is taken of this first page by among other things including results from different sources (e.g., news, images, and videos).

Evaluations of search logs show that between 60% and 85% of searchers only view the first search result page [4, 7]. In only 4.3% of the sessions users look at more than 3 search result pages for a query. So, for most users and search tasks, the popular search engines are adequate. However, this still leaves searches for which more than only the first few result pages matter. Therefore, we have chosen to take a rather controversial stance and look at how recall-oriented search tasks on the Web could be better supported.

ESAIR'13, October 28, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2513204.2513219 ...$15.00.

Two typical examples of recall-oriented search tasks are brand monitoring and e-discovery/forensic investigations. When monitoring a brand, a company wants to retrieve all relevant user generated content on the Web that mentions or gives an opinion about the brand. While most users will be satisfied with the company's homepage when searching for a brand, the brand monitoring task might have the same query terms as starting point, but its goals are entirely different. Forensic investigation is another example of a recall-oriented task. Any detail on any web page can be important in solving an investigation. Most likely, crucial information is even absent on the 10 most popular pages about the topic of investigation (e.g., a person or an event). So, an exhaustive search is needed to be able to generate an accurate and optimized image. A E-discovery is an active field of research because of its usecase in the United States, where e-discovery refers to the requirement that documents and information in electronic form stored in corporate systems be produced as evidence in litigation [6].

In this article we focus on forensic investigation and propose a recall-oriented search system. In the next section, we discuss the characteristics of recall oriented search tasks and strategies to improve recall. In Section 3, semantic annotations, domain adaptation, and search results in context are discussed. Last, in Section 4, we present our conclusions.

2. RECALL-ORIENTED SEARCH

The following characteristics can be attributed to recall-oriented search tasks on the Web:

- Often searches will focus on user generated content. Huge amounts of textual data are generated every day in social networks, blogs, and forums and via tweets. In contrast, more focused and high quality content can be found on (official) homepages.
- The typical 2 to 3 word query used to search the Web is not sufficient to ensure the recall of all documents relevant to your search topic. Issuing multiple search queries containing different additions and replacements of search words increases the recall of relevant documents.
- Queries do not have to be answered instantly. A user is likely to spend a considerable amount of time on the search results, so he is willing to wait a little to get good results back.

In this article, we want to investigate how we can exploit existing web search engines for recall oriented search. So, we will exploit

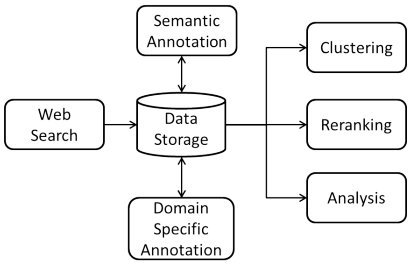

Figure 1: High level system architecture.

Table 1: The number of annotations on the analyzed corpus of news data.

	locations	persons	organizations
# distinct instances	7,809	15,303	20,115
# assignments	108,805	90,559	133,285

their strengths and devise workarounds for their weaknesses. The alternative would be to develop a general Web search engine from scratch, which would be optimized for recall oriented search. This would be the ultimate test for the proposed procedure. However, the development of such a general Web search engine from scratch beyond the scope of this article. Nevertheless, especially for search in a specific domain, this should be acknowledged as a feasible option.

The first issue to deal with is: how to retrieve a larger number relevant results? A 2 to 3 word keyword query is not sufficient. The query should be expanded either by the user or via the use of search and domain context [3]. In practice, sending multiple search queries to the search engine will lead to more results than adding terms to the query, even when the OR operator is used, since most search engines only return the first 1,000 search results of any search.

Commonly, recall-oriented search engines are optimized for a certain application domain. The context of the search is known and, hence, can be exploited to increase the number of search results retrieved. Multiple search queries can be created by using synonym lists or adding general terms related to the domain. Alternatively, recall can be increased via the inclusion of multiple search engines for the search at hand. With popular sets of query words, major searches engines will provide similar results. However, the further the results are explored, the bigger the differences between the search results will become. Also with less common sets of query words, the results will show larger deviations.

3. SEARCH RESULTS IN CONTEXT

Figure 1 shows our high-level system architecture. The system uses Bing and Google's search APIs to collect Web search results. The URLs are scraped and, subsequently, the pages' textual content is extracted and saved in a database. All text is annotated and, hence, is ready to be clustered and reranked.

In recall-oriented search, it is still important to find a good balance between the precision and the recall of the search results. When too many irrelevant results are found, the data will become too noisy to drive conclusions from and, consequently, the user's motivation to go through the search results will decline. Additionally to the match of the search words, other criteria for relevance can be considered [10].

Semantic annotations can help in finding a balance between the precision and the recall of the search results. They can significantly contribute to reranking and filtering out search results, which are not relevant in the particular search context. Adding semantic annotations to all search results can help the user to apply a divide and conquer strategy to process the search results. Semantic annotations can be used to cluster or rerank the search results into many dimensions. Using clustering larger numbers of search results can be processed more quickly by removing irrelevant clusters from your search results, and zooming in on interesting clusters. There are many types of semantic annotations that can be made, here we focus on temporal annotations and entity type annotations (i.e., persons, locations and organization).

3.1 Annotations

Named entity recognition is a well studied problem (see [5] for an overview). Here, in particular, we consider how to exploit the entity tags to improve recall-oriented search. We analyzed a corpus containing 86,024 news items collected in 2012 and 2013. They have been annotated using the LingPipe tool kit[1]. These annotations were originally created for the EU FP7 project Virtuoso [2].

The occurrences of locations, persons, and organizations in our news corpus follow a log distribution, as is common for Web-based corpora. The majority of locations, persons, and organizations are only mentioned once or a few times, while only a few of them are mentioned frequently. In Figure 2, the occurrences of locations, persons, and organizations in the annotated corpus of news data is presented. The absolute number of occurrences per annotation type is plotted on a logarithmic scale. Table 1 provides the number of distinct instances as well as the total number of assignments per annotation type.

Search results can usually be filtered using the date and time a page was last updated. However, many more date indications can be found on the web pages itself. For example, in a forum each post will have a date and time attached to it. Annotations tools such as Heideltime can extract these date and time stamps[8].

3.2 Clustering and Reranking

Here, we describe two methods to use the annotated search results: clustering, and reranking. The first method is to cluster the search results is to make clusters for all entities and dates found. Then, a document can be assigned to multiple clusters, incase it has multiple distinct annotated entities or dates. However, as is illustrated in Table 1, it is likely that too many clusters will be formed. Many of these clusters would containing only one or a few documents. This issue could be solved by clustering the entities and dates as well.

The second method is to rerank search results; for example, by date. That is, the closer the extracted dates from the Web pages are to a certain target date specified by the user, the higher a result will be ranked. Distances between dates can be conveniently calculated. However, distances between locations, organizations, and persons are more challenging to calculate. As an initial rule of thumb, the system can suggest frequently occurring entities and/or dates as target entity and/or data.

For clustering as well as reranking, distances between annotations need to be calculated. There are several options:

- Use co-occurrence statistics from the total set of search results [2].

[1] `http://alias-i.com/lingpipe` [Last accessed: July 18, 2013]

[2] `http://www.virtuoso.eu/` [Last accessed: July 18, 2013]

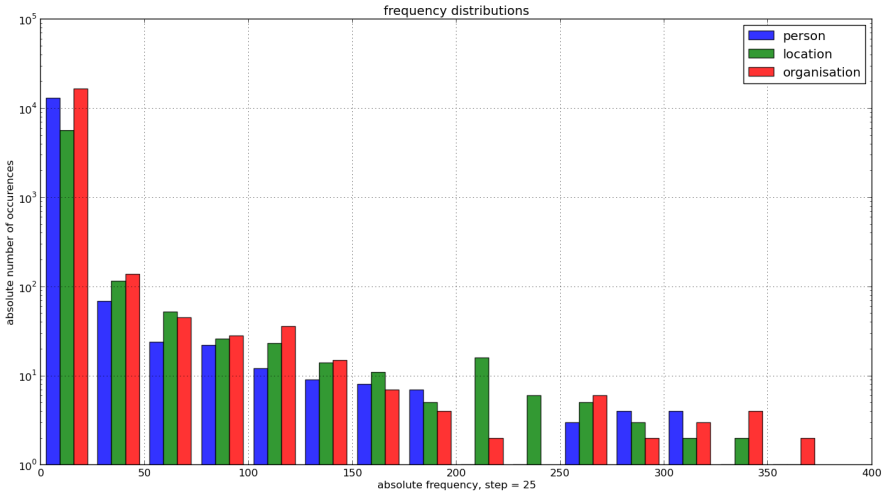

Figure 2: Occurrences of semantic annotations: locations, persons, and organizations.

- Exploit an external corpus like Wikipedia to determine the distance between entities [9, 10].

- In the case of locations, locations can be mapped to geographical locations. Subsequently, the distance between these geographical locations can be conveniently calculated. This process can be facilitated using an external corpus such as GeoNames[3]. In addition, relations like city *x* lies in country *y* can help to determine distances between locations on different scales.

- In the case of persons, search results often either come from social networks or can be related to them. Then, it is possible to conduct a social network analysis. Such analysis can be used to calculate several distance measures [1].

3.3 Domain Adaptation

To efficiently perform search gains can be achieved by adapting systems to their application domains. Search in a specific domain can benefit from exploiting the domain context (e.g., forensic investigation). In contrast to general Web search, the most relevant results for this domain are usually not the most popular results. One of the features that is an indication for a relevant result in the forensic search domain is the use of aggressive language. If a Web page contains a lot of offensive words, this increases the chance that this result will be relevant.

For a specific use case such as monitoring possible riots related to soccer games, a tailored vocabulary can be defined. Such a vocabulary can be either defined manually by forensic experts or it can be derived from documents tagged as relevant to the use case. As mentioned in Section 2, domain context can be used for query expansion. Multiple search queries can be generated by combining the search keywords with domain keywords. Although there will be overlap in the search results of the different queries, overall recall will be boosted. Also, the domain specific context annotations can be leveraged in the same way as the semantic annotations discussed before: to cluster or rerank the complete set of results.

[3] `http://www.geonames.org/` [Last accessed: July 18, 2013]

4. CONCLUSION

In this article, we have proposed several ways in which semantic annotations can provide support for recall-oriented search on the Web. This is work in progress. We have planned to implement a full recall-oriented search system and evaluate its performance and efficiency with users from the domain of forensic investigation. Questions we aim to answer are: 'How to increase recall when searching the Web?' and 'How to use context and/or annotations to help users process large numbers of search results more efficiently?'

5. REFERENCES

[1] L. Adamic and E. Adar. How to search a social network. *Social Networks*, 27(3):187–203, 2005.
[2] I. Dagan, L. Lee, and F. C. Pereira. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1-3):43–69, 1999.
[3] G. Fischer. Context-aware systems: The 'right' information, at the 'right' time, in the 'right' place, in the 'right' way, to the 'right' person. In *Proceedings of the Conference on Advanced Visual Interfaces (AVI 2012)*, pages pp. 287–294, 2012.
[4] B. J. Jansen and A. Spink. How are we searching the world wide web? a comparison of nine search engine transaction logs. *Information Processing & Management*, 42(1):248–263, 2006.
[5] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.
[6] D. Oard and W. Webber. Information retrieval for e-discovery. *Foundations and Trends in Information Retrieval*, 7(2–3):pp. 99–237, 2013.
[7] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. Analysis of a very large web search engine query log. In *ACM SIGIR Forum*, volume 33, pages 6–12. ACM, 1999.
[8] J. Strötgen and M. Gertz. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 2012.
[9] M. Strube and S. P. Ponzetto. WikiRelate! computing semantic relatedness using Wikipedia. In *AAAI*, volume 6, pages 1419–1424, 2006.
[10] F. van der Sluis, E. L. van den Broek, R. J. Glassey, E. M. A. G. van Dijk, and F. M. G. de Jong. When complexity becomes interesting. *Journal of the American Society for Information Science and Technology*, [in press], 2014.