NewsSearch: Search and Dynamic Re-ranking over News Corpora

Luka Stopar, Blaž Fortuna, Marko Grobelnik
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 477 3933
e-mail: {luka.stopar, blaz.fortuna, marko.grobelnik}@ijs.si

ABSTRACT

NewsSearch proposed in this paper is a search engine interface, which allows users to visually re-rank search results. When a query is first made, the results are visualized using topic, concept and location widgets on the client side. The user is then able to filter and re-rank complete results by interacting with the visualization widgets. Filtering and re-ranking is performed on the server side, allowing the user to seamlessly browse through millions of results.

1 INTRODUCTION

Search engines emerged as useful tools for information seeking, allowing users to access relevant information quickly through simple keywords queries. Knowing how to use them can save time and get reliable results.

Search engines today have to display tens of thousands of ranked documents, which lead to user paralysis and poor choices. A typical query easily results in thousands or more results. However, on average less than 6% of users would click on a link from the second page of results [6]. This leaves a large unexplored space of results, which are hard to comprehend, since it requires manually iterating through results pages. NewsSearch [7] addresses this issue by visualizing the long tail of search results. The documents are grouped into related groups according to several criteria, allowing the user to filter and re-rank the documents based on the similarity to each group.

This paper improves upon the work from [7] in the following ways. First, re-ranking is performed on the server side, which increases the amount of results, which can be handled with this approach. Second, query expansion is made pare of re-ranking, further increasing the scalability of the approach. Finally, new visualization widget was introduced, allowing re-ranking and filtering according to geospatial information, associated with the document.

The paper is structured as follows. We start by giving an overview of news collection and indexing services, which are used in the paper. This is followed by introducing different types of visualization widgets, which are used to visualize the search results. We conclude by presenting the query extension based re-ranking approach and the system prototype.

2 NEWS COLLECTION AND INDEXING

NewsSearch uses news collection and indexing services developed at J. Stefan Institute (JSI) and available at http://newsfeed.ijs.si [1]. The services provide a real-time aggregated stream of news articles by crawling RSS-enabled news providers across the world. The crawler currently downloads 50,000-100,000 articles per day from about 100,000 RSS feeds. The current archive contains about 25 million articles and begins in May 2008. The stream of articles is serialized into XML and segmented by time into compressed files with several megabytes in size.

The news collection service collects articles from a multitude of languages, however in this work we only focused on a subset of English news sources. Each article is processed using text enrichment service Enrycher [8], which extracts topics and entities from the articles. Additionally, for all major news sources, additional meta data was collected, including their location.

For indexing and search the news collection we use News Miner, which is a system for processing and indexing news corpora and is entirely based on the code base developed at JSI. Each article is indexed across several dimensions (facets) using an inverted index. The system allows for retrieval of any Boolean combination of the facets.

Processing of a news feed contains the following steps each being executed by a separate process:

1. Retrieve the articles from the news feed (e.g. using News Collection service's Python script)
2. Parse the article and prepare the fields for indexing (e.g. tokenizing text)
3. Add the article to the index

Once the article is indexed, it can be accessed through keyword queries.

3 VISUALIZATION WIDGETS

NewsSearch interacts with the user through a series of widgets, which are designed to visualize different dimensions of the complete search results collection.

The user starts the search by entering a search query and the widgets are generated based on the resulting documents. The system starts by using default search engine ranking (e.g. BM25 [9]). The user can influence the ranking by moving the target with a mouse to different parts of widgets, which causes re-ranking of the results.

3.1 WORLD WIDGET

The world widget, shown in Figure 1, enables the user to reorder results based on the locations assigned to the document. This can be either the locations mentioned in the article, or the location of the news outlet, which produced the article. When the user moves the target, the coordinates are sent to the server and the results weighted and reordered based on the distance from the target coordinates.

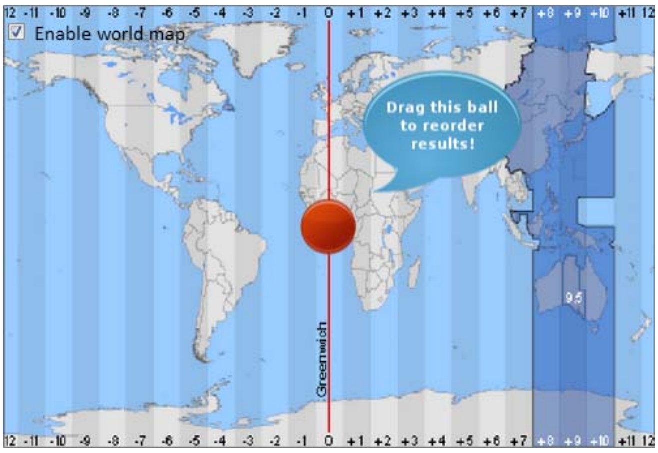


Figure 1. The world widget enables the user to re-rank results based on locations assigned to the documents.

3.2 TOPIC WIDGET

The topic widget, shown on the left in Figure 2, allows the user to filter and re-rank results based on the topics they want to view.

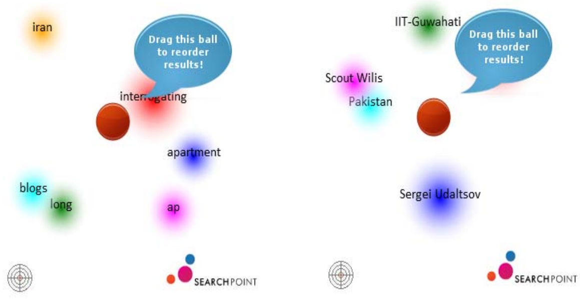


Figure 2. Topic and concept widgets enable the user to filter and re-rank results, based on keywords and concepts.

The widget displays clusters of articles represented by the most relevant keywords in each cluster. When the user moves the target, the new coordinates are sent to the server, which first computes the distance to each cluster and extracts the most representative keywords for particular coordinates. These keywords are then used to extend the query issued by the user, and sent to the search engine to filter non-relevant results and rank higher the remaining results using BM25 weights.

To construct the clusters, NewsSearch represents documents as vectors with the standard Bag-of-words representation where there is a dimension for each word. It then uses the K-Means++ [3, 4] algorithm, with cosine similarity, to cluster the documents into several clusters, which are displayed by the widget. Cluster centroid vectors are used to extract the most representative keywords for each cluster.

To visualize the clusters the keyword widget uses Multidimensional scaling (MDS [5]) to embed the centroids onto a two dimensional plane, similar as in [2].

3.3 CONCEPT WIDGET

The concept widget, shown on the right in Figure 2, allows the user to filter and reorder results based on the concepts contained in the articles. Concepts are extracted from the articles using the Enrycher service, and ranked using the News Miner service.

The widget displays clusters of articles represented by the most relevant concepts in each cluster. Like with the keyword widget, when the server receives new coordinates it computes the distance to each cluster and extracts the most representative concepts. It then constructs a new query using these concepts, to filter non-relevant ones and rank the remaining results using BM25 weights.

Unlike the topic widget, the concept widget clusters concepts and not documents. Each concept is represented as a “Bag-Of-Documents”, where each document represents one dimension and a coordinate represents the number of times a concept appears in that document. The widget then uses the K-Means++ [3] algorithm, with cosine metrics.

Like the topic widget, the concept widget also uses MDS to visualize clusters on a plane.

3.4 DMOZ WIDGET

The DMOZ widget, shown in Figure 3, allows the user to reorder results based on their DMOZ categorization, provided by the Enrycher service.

The widget displays a tree of categories, extracted from the results, where the size of a category depends on the number of articles it contains.

When the server receives new coordinates, it reorders results based on the distance to their category.

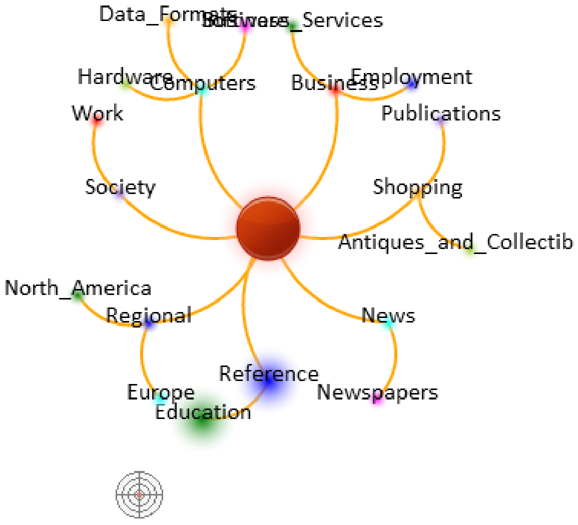


Figure 3. The DMOZ widget enables the user to reorder results based on their DMOZ category.

4 RE-RANKING

When re-ranking is performed, the server side is passed information about the position of all the targets in the widgets. The server constructs a new query by taking the user query, and extending it by weighting keywords and/or concepts contained in nearby clusters and selecting the top 5.

Keyword and/or concept weight is computed with the following formula:

$$w = w_c w_k.$$

The weight of the cluster w_c, is the same for all keywords and/or concepts in a cluster, and is computed using the Gaussian kernel

$$w_c(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}.$$

The weight w_k of a keyword and/or concept inside the cluster is computed using the position i of the keyword and/or concept in the cluster's centroid using following formula:

$$w_k(i) = e^{-\frac{i}{2}}$$

The new query is generated by conjucting a disjunction of the selected keywords/concepts. For example, let A be the original search term and let B, C, D be the keywords selected by the weighting procedure the new query Q will then be:

$$Q = A \wedge (B \vee C \vee D)$$

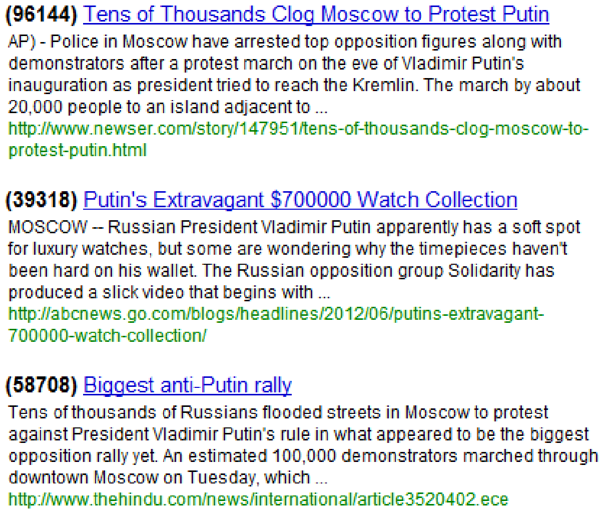
When the query results are obtained, they are weighted and ranked. The weight is computed using TF-IDF multiplied by the weight of the keyword/concept, so with the same TF-IDF keyword B will have a higher weight then keyword C. The world widget weight is computed using a Gaussian kernel. All three weights are multiplied and the results are sorted.

The results are then passed to the client side and displayed in a standard way.

5 EXAMPLE

In an example, we will create a query on a well-known Russian politician Vladimir Putin.

When the query is performed, the results are weighted using just the TF-IDF weight of keyword "putin". The results are shown in Figure 4.

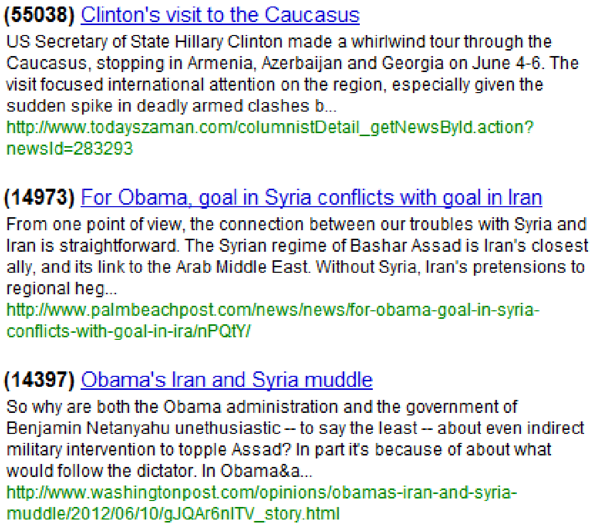


(96144) Tens of Thousands Clog Moscow to Protest Putin
AP) - Police in Moscow have arrested top opposition figures along with demonstrators after a protest march on the eve of Vladimir Putin's inauguration as president tried to reach the Kremlin. The march by about 20,000 people to an island adjacent to ...
http://www.newser.com/story/147951/tens-of-thousands-clog-moscow-to-protest-putin.html

(39318) Putin's Extravagant $700000 Watch Collection
MOSCOW -- Russian President Vladimir Putin apparently has a soft spot for luxury watches, but some are wondering why the timepieces haven't been hard on his wallet. The Russian opposition group Solidarity has produced a slick video that begins with ...
http://abcnews.go.com/blogs/headlines/2012/06/putins-extravagant-700000-watch-collection/

(58708) Biggest anti-Putin rally
Tens of thousands of Russians flooded streets in Moscow to protest against President Vladimir Putin's rule in what appeared to be the biggest opposition rally yet. An estimated 100,000 demonstrators marched through downtown Moscow on Tuesday, which ...
http://www.thehindu.com/news/international/article3520402.ece

Figure 4. First three search results for query "putin".



(55038) Clinton's visit to the Caucasus
US Secretary of State Hillary Clinton made a whirlwind tour through the Caucasus, stopping in Armenia, Azerbaijan and Georgia on June 4-6. The visit focused international attention on the region, especially given the sudden spike in deadly armed clashes b...
http://www.todayszaman.com/columnistDetail_getNewsById.action?newsId=283293

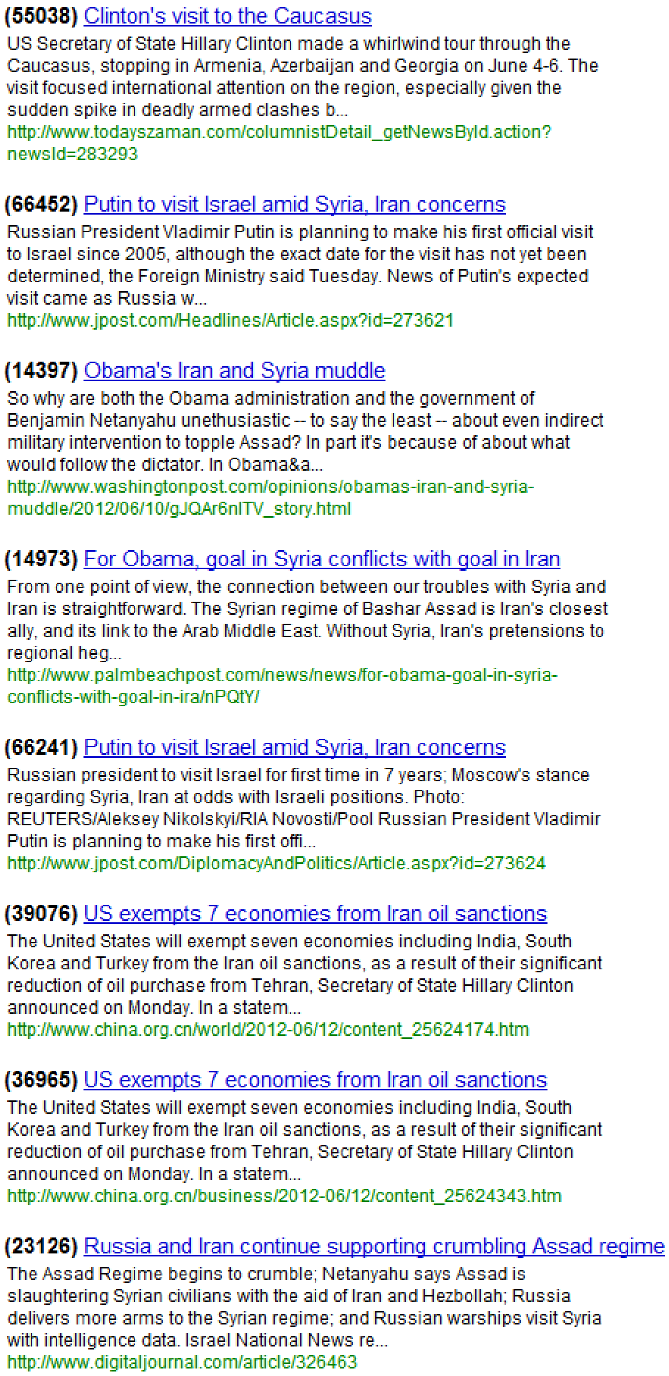
(14973) For Obama, goal in Syria conflicts with goal in Iran
From one point of view, the connection between our troubles with Syria and Iran is straightforward. The Syrian regime of Bashar Assad is Iran's closest ally, and its link to the Arab Middle East. Without Syria, Iran's pretensions to regional heg...
http://www.palmbeachpost.com/news/news/for-obama-goal-in-syria-conflicts-with-goal-in-ira/nPQtY/

(14397) Obama's Iran and Syria muddle
So why are both the Obama administration and the government of Benjamin Netanyahu unethusiastic -- to say the least -- about even indirect military intervention to topple Assad? In part it's because of about what would follow the dictator. In Obama&a...
http://www.washingtonpost.com/opinions/obamas-iran-and-syria-muddle/2012/06/10/gJQAr6nITV_story.html

Figure 5. First three search results when the target is moved to topic "Iran".

When the target on the keyword widget is moved on to the "Iran" cluster, a new query is constructed on keywords "putin", "Iran", "Syria" and "Visit", and the results are weighted using TF-IDF giving keyword "Iran" the largest initial weight. The results are shown in Figure 5.

Then the target is moved to concept “Regional”. A query is constructed on keywords “putin”, “Iran”, “Syria” and “Visit”, and on concepts “Regional”, “Syria”, “Society and Culture”, “Damascus” and “United States”. The keywords are weighted same as before and the weight is multiplied by the weight of the concepts which are also weighted with TF-IDF, yielding results shown in Figure 6.



(55038) Clinton's visit to the Caucasus
US Secretary of State Hillary Clinton made a whirlwind tour through the Caucasus, stopping in Armenia, Azerbaijan and Georgia on June 4-6. The visit focused international attention on the region, especially given the sudden spike in deadly armed clashes b...
http://www.todayszaman.com/columnistDetail_getNewsById.action?newsId=283293

(66452) Putin to visit Israel amid Syria, Iran concerns
Russian President Vladimir Putin is planning to make his first official visit to Israel since 2005, although the exact date for the visit has not yet been determined, the Foreign Ministry said Tuesday. News of Putin's expected visit came as Russia w...
http://www.jpost.com/Headlines/Article.aspx?id=273621

(14397) Obama's Iran and Syria muddle
So why are both the Obama administration and the government of Benjamin Netanyahu unethusiastic -- to say the least -- about even indirect military intervention to topple Assad? In part it's because of about what would follow the dictator. In Obama&a...
http://www.washingtonpost.com/opinions/obamas-iran-and-syria-muddle/2012/06/10/gJQAr6nlTV_story.html

(14973) For Obama, goal in Syria conflicts with goal in Iran
From one point of view, the connection between our troubles with Syria and Iran is straightforward. The Syrian regime of Bashar Assad is Iran's closest ally, and its link to the Arab Middle East. Without Syria, Iran's pretensions to regional heg...
http://www.palmbeachpost.com/news/news/for-obama-goal-in-syria-conflicts-with-goal-in-ira/nPQtY/

(66241) Putin to visit Israel amid Syria, Iran concerns
Russian president to visit Israel for first time in 7 years; Moscow's stance regarding Syria, Iran at odds with Israeli positions. Photo: REUTERS/Aleksey Nikolskyi/RIA Novosti/Pool Russian President Vladimir Putin is planning to make his first offi...
http://www.jpost.com/DiplomacyAndPolitics/Article.aspx?id=273624

(39076) US exempts 7 economies from Iran oil sanctions
The United States will exempt seven economies including India, South Korea and Turkey from the Iran oil sanctions, as a result of their significant reduction of oil purchase from Tehran, Secretary of State Hillary Clinton announced on Monday. In a statem...
http://www.china.org.cn/world/2012-06/12/content_25624174.htm

(36965) US exempts 7 economies from Iran oil sanctions
The United States will exempt seven economies including India, South Korea and Turkey from the Iran oil sanctions, as a result of their significant reduction of oil purchase from Tehran, Secretary of State Hillary Clinton announced on Monday. In a statem...
http://www.china.org.cn/business/2012-06/12/content_25624343.htm

(23126) Russia and Iran continue supporting crumbling Assad regime
The Assad Regime begins to crumble; Netanyahu says Assad is slaughtering Syrian civilians with the aid of Iran and Hezbollah; Russia delivers more arms to the Syrian regime; and Russian warships visit Syria with intelligence data. Israel National News re...
http://www.digitaljournal.com/article/326463

Figure 6. Search results when the topic widget is set to “Iran” and the concept widget is set to “Regional”.

Finally when the target on the world widget is moved over New York, the previous weight is multiplied with the world position weight producing the final result.

6 CONCLUSIONS

NewsSearch aims to enable the user to more efficiently browse search results, which increases user productivity and the quality of information gathered. It can be used for search verticals or enterprise search, when one cannot invest a lot of money in tuning the ranking.

A demo can be found at http://searchpoint.ijs.si/ and where the data source used is Microsoft’s search engine Bing.

7 ACKNOWLEDGEMENTS

This work was supported by the Slovenian Research Agency, the ICT Programme of the EC under PASCAL2 (IST-NoE-216886) and XLike ((FP7 -ICT-288342-STREP).

References

[1] Trampuš, M.; Novak, B. Internals of an aggregated web news feed. In: Proceedings of the fifteenth international multiconference Information Society 2012. Ljubljana: Institut Jožef Stefan, 2012.

[2] Fortuna, B.; Mladenić, D.; Grobelnik, M. Visualization of text document corpus. Informatica 29 (2006).

[3] Arthur, D.; Vassilvitskii, S. k-means++: the advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms (2007).

[4] Jain, A. K.; Murty, M. N.; Flynn, P. J. Data clustering: a review. ACM Computing Surveys 31, 3 (1999).

[5] Carroll , J. D.; Arabie, P. Multidimensional scaling. In M.R. Rosenzweig and L.W. Porter (Eds.): Annual Review of Psychology 31 (1980).

[6] http://insights.chitika.com/2010/the-value-of-google-result-positioning/

[7] Pajntar, B.; Grobelnik. M. SearchPoint – a New Paradigm of Web Search. 17th International World Wide Web Conference (WWW2008) Developers Track.

[8] Štajner, T.; Rusu, D.; Dali, L.; Fortuna, B.; Mladenić, D.; Grobelnik, M. A service oriented framework for natural language text enrichment. Informatica (Ljublj.), 2010, vol. 34, no. 3, 307-313.

[9] Manning, C.D.; Schutze, H. Foundations of statistical Natural Language Processing (MIT Press, 1999).