Multilingual Trend Detection in the Web*

Jan Stutzki

Universität der Bundeswehr München
Werner-Heisenberg-Weg 39, Germany
jan.stutzki@unibw.de

Abstract

This paper represents results from our ongoing research project in the foresight area. The goal of the project is to develop web based tools which automatically detect activity and trends regarding given keywords. This knowledge can be used to enable decision makers to react proactively to arising challenges.

As for now we can detect trends worldwide in more than 60 languages and assign these trends accordingly to over 100 national states. To reach this goal we utilize the big search engines as their core competence is to determine the relevance of a document regarding the search query. The search engines allows slicing of the results by language and country.

In the next step we download some of the proposed documents for analysis. Because of the amount of information required we reach the field of Big Data. Therefore an extra effort is made to ensure scalability of the application.

We introduce a new approach to activity and trend detection by combining the data collection and detection methods. To finally detect trends in the gathered data we use data mining methods which allow us to be independent from the language a document is written in. The input of these methods is the text data of the downloaded documents and a specially prepared index structure containing meta data and various other information which accumulate during the collection of the documents.

We show that we can reliably detect trends and activities in highly active topics and discuss future research.

1998 ACM Subject Classification H.3.5 Web-based services

Keywords and phrases Information Retrieval, Web Mining, Trend Detection

Digital Object Identifier 10.4230/OASIcs.SCOR.2014.16

1 Introduction

The Web is the number one information source of our time. Almost all the knowledge of todays world is available online. Current news gets published almost in real time. Developments of events and what people are interested in is mirrored by the content of popular and relevant online documents. For a long time now the Web is in an transition from a tool used by scientists to an everyday commodity used by anyone [1]. While most of the web pages are still written in English the part which is not is growing in absolute and relative size [2].

With the global community expressing its interests and worries online (see Figure 1) it seems appropriate to develop a system which is able to track certain topics and their developments trough published Web documents. It is required to have an reliable detection mechanism for activities and trends available to be able to focus on certain activities and keep track of them over time.

* This work was partially supported by the Planungsamt der Bundeswehr.

4th Student Conference on Operational Research (SCOR'14).
Editors: Pedro Crespo Del Granado, Martim Joyce-Moniz, and Stefan Ravizza; pp. 16–24

OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

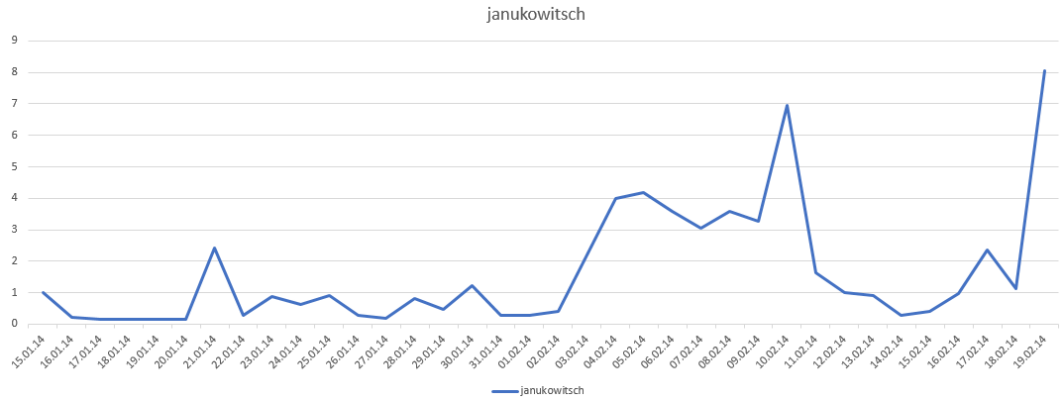

Figure 1 Relative term frequency of "janukowitsch" at the brink of the Ukraine unrest.

Our work focuses on activity and trend detection in the Web. We develop a concept which enables the collection and analysis of potentially relevant data in a scalable and robust way. We propose several methods which are the focus of this paper for automated activity and trend detection in the Web using web services to do the relevance rating for us. We analyze these methods in regard of their ability to detect either and check their performance for different settings. The setting differs in focus of the application and availability of data. To evaluate the methods and as a proof of the concept we developed a prototype (see Figure 2). The web based application is easy to use and provides quick access to the results of most of the proposed methods.

The paper is structured as follows. In Section 2 we explain how data is acquired and what kind of information is collected in what fashion and explain our overall concept for a new approach to activity and trend detection. Section 3 discusses the definitions of trends and activity and introduces the methods which we propose for detection and tracking. The analysis for each method covers its strengths and weaknesses. The applicability of the method for trend detection and activity measurement is also covered. Section 4 contains the conclusion and gives an outlook for further research.

2 Concept

This section focuses on what information is required by the user of the proposed system and which information is gathered automatically. Our concept for data retrieval and analysis consists of six steps:

1. defining search terms
2. translating search terms
3. selection country and language combinations
4. querying search engines for documents
5. downloading the documents
6. analyzing the documents

The task of presenting the results of the documents is regarded as an extra function which is independent of the data collection and analysis process. The idea behind our concept is that experts define the general area in which to look for trends. An expert might be tasked to find new trends and/or centers of activity and developments for a certain topic. Without any further knowledge it may be hard to know where to start.

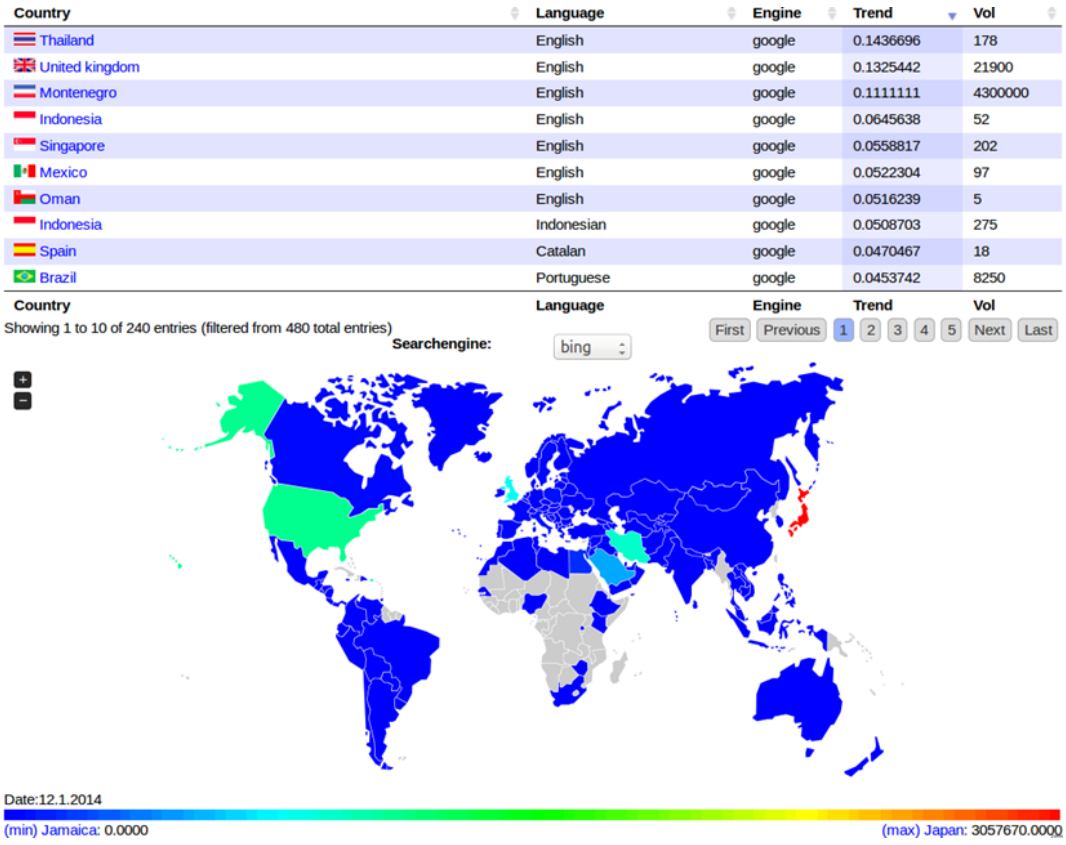

Figure 2 Screen shot of the prototype.

The expert defines a few topic specific search terms in his or her own language and uses a web interface to transfer them to a web based tool. The tool can automatically translate the search terms in the most used languages and allows the user to adapt the automatic translation in case of mistakes [5].

In the next step the user selects which language should be considered for which country. This saves a lot of resources as not all combinations are required. By default for each country the official languages (if available) and English will be preselected as almost 70 % of the Web is in English [3] and English therefore can be assumed as an universally used language independent of the country.

Using the county and language combinations together with the translated search terms the system starts to query one or more search engines for URLs of relevant documents. As Google and Bing do punish scraping [7] we use the official APIs which for Google is Custom Search Engine (CSE) and Bing Azure Websearch. Scraping retrieves the resulting URLs from the official HTML result pages of the search engine by simulating a user and a browser via software. Web search engines disallow the use of scraping because software does not click on advertisements or leaves personal information so there is no source for revenue for the service provider. The punishment for scraping is often done by excluding the client from the service which would be detrimental to our goal of retrieving URLs in fixed intervals. We refer to one run of downloading documents as one crawl and the component doing the downloading the crawler.

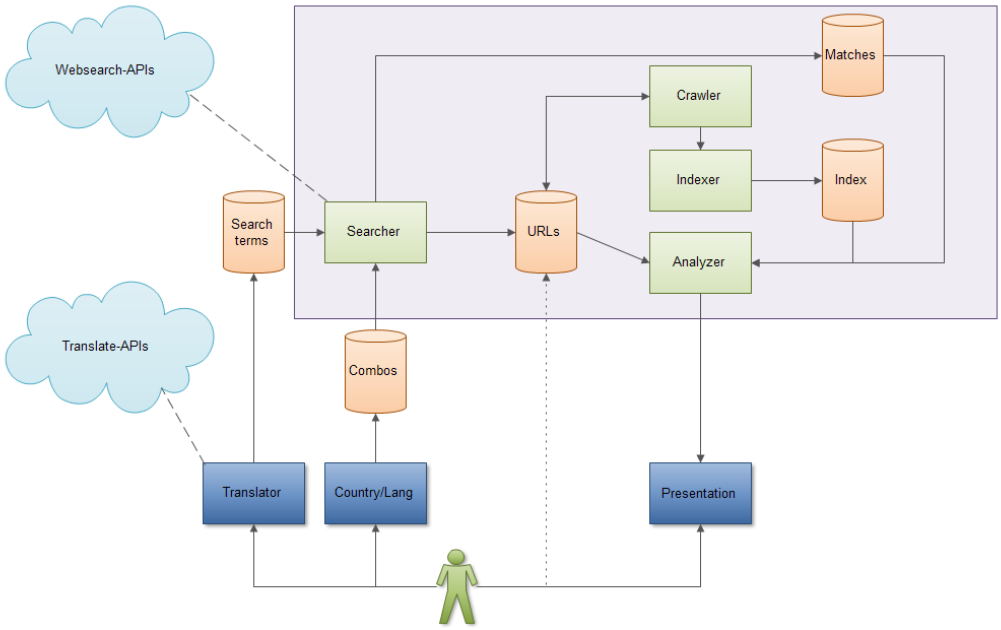

Figure 3 System structure.

The fixed interval between crawls is termed the crawl interval. The shorter the crawl interval is, the less time the data has to change. For very static topics a long search interval might be sufficient while for active topics like news a daily search interval is already at risk of loosing developments. For the analysis of the results of the crawls it is important that the interval between crawls is equidistant.

The crawler will then start to download the documents indicated by the URLs from the search engines. Not all documents are available to any crawler [4] and not every document is actually usable for trend or activity detection. For this reason the crawler has to be exceptionally robust. For analysis we have several information sources. We have the data from the search engines which includes a list of URLs and an estimate of overall matches to the query. We also have the results of multiple crawls and the time when each crawl was done. From the crawls we can deduct several data structures we can use for further analysis. The system is shown in Figure 3. The whole violet box is operating unsupervised and is responsible to perform the crawls in the predefined intervals.

3 Methods

We analyzed several of the following methods and evaluated them in regards to their ability for activity measurement and trend detection:

- Relative Term Frequency
- Estimated Matches
- Page Updates
- New Sources

All methods have in common that results can only be seen relative to previous results. Informally we define trend as a directed activity. An activity is a change between t_0 and t_1 which is detectable within the data we collect. We are particularly interested in activity created by humans. So our methods have to filter the changes in the data for noise and

for changes actually resulting from human behavior. With this trend definition everything done by humans would qualify as a trend. The direction of an activity can be deducted by an increase or decrease of an inspected parameter. While there is no problem with seeing everything as a trend we use "reach" as a property of a trend which allows us to filter for trends which occur on a given percentage of documents. One objective unit for a trend we propose is a term as terms pose a standard used by most documents.
Activity is more generally the deviation between two measurements. This definition allows us to utilize all activity detection methods to some degree for trend detection as a trend is a directed activity. On the other hand activity measurement methods do not necessarily allow us to detect trends as the direction (e. g. rising/falling) component might be missing or carries no value. We use the activity to measure if a topic is still active e. g. there is still research done or has become static.

During our research we also considered synonyms and stemming to be relevant for trend detection as we develop some methods which are text based. We did not regard synonyms as they are heavily context dependent and our index structure currently does not support this. This might change with a positional index but the challenge to extract or collapse the right synonyms for the supported languages remains. Also there are few sources for comprehensive data about multilingual synonyms. After experiments with stemming we decided against it as too much relevant information is lost during stemming (e. g. gender for job descriptions in German) so that a trend might get lost among the other terms which share the same stemmed form while only slightly reducing the dictionary size of the inspected languages. Experiments with a German dictionary and various sources lead to an approximate reduction of the dictionary by only 15 %. With the same reasoning we decided against the usage of character folding. While it is certainly reasonable to use it in a information retrieval context we concluded that its effects are detrimental to trend detection.

For each method we evaluate what data is required and if it is available. Furthermore we consider the dimensions inspected by the methods and a fitting way of visualization and estimate how complex a method is to run on current standard hardware (Intel i5, 8 GB RAM). Because we target documents in several languages and therefore our methods have to work with a multitude of languages we disregard semantic approaches which are usually language specific and focus on statistical analysis of the text corpus at hand.

3.1 Relative Term Frequency

We define the trend of a term t_w as the incline of the trend line of data points of the relative term frequency $F = (f(w)_0, \ldots, f(w)_i)$. The term w has to be part of each data point d_i inspected. The relative frequency $f(w)_i$ is calculated $f(w_i) = |w_i|/|w_0|$ using the frequency of t_0 as reference. The trend line is fitted with least square linear regression to the whole time series inspected. Using relative term frequency allows us to work without stop lists.

This method requires a time frame as an additional input as terms might not be present in the documents crawled at a given time. While the absence of a term could be used as an indicator it saves memory and improves the performance when terms which are not present from t_0 to t_i are ignored.

The relative frequency can be used as an indicator for activity as any change in the relative frequency can be tracked back to change in the underlying text. A high activity is expressed by huge changes in the overall relative term frequency (see Figure 1). To get good results it is necessary to look at several, if possible, all terms.

For trend detection this method is recommendable as each relative term frequency time line can be understood as a trend indicator. Paired with other metrics as for example "reach"

it can indicate which terms are currently rising over proportionately in frequency which we interpret as interest of the humans behind the machines.

In combination with least square linear regression a long term trend can be derived from the data and enable the user to focus on a few exceptional instances of terms for further inspection. Because this method needs to tokenize a text in order to extract terms, problems with languages which do not have an explicit word delimiter arise (e. g. Chinese and Japanese). For other languages this method exceeds expectations and will be part of future research to improve the trend filter and make exceptional trends more obvious. The problem with the tokenizer is an area of active research [6].

We can use this method for the dimensions: language, country, time and text corpus. Depending on the flexibility of the time frame and the inspected dimensions the computational complexity of the method is rather high as the tokenizer needs preprocessed input where the documents are sanitized and cleaned of any markup. While this can be precalculated if the time frame is fixed, non fixed time frames need to be calculated on demand. This can be supported by index structures which are generated after a crawl is finished, but for this to be feasible the dimensions and aggregation level needs to be defined in advance.

Visualization is done on a per term basis. For each term a graph is generated for the inspected time frame. The terms are presented as a sortable list. The list is sortable by term, reach and incline of the trend line.

3.2 Estimated Matches

The number of "Estimated Matches" is acquired during a crawl. It is the number of estimated results as returned by a search engine as part of the reply to a specific query. As all queries are country and language specific we can assign the number of estimated matches to a country and a language.

Depending on the target variable, the results of one crawl may be sufficient. Questions like "where in the world are the most matching documents located" can be answered. Because all languages of a country are treated separately we aggregate the number of estimated documents for each country over all corresponding languages. If a trend is to be detected, at least two crawls are required to retrieve two data points which then can be used to estimate a trend. The estimate number of matches for a given query provided by the search engines varies (see Figure 4). Especially Bing offers unreliable results. In order to give an informed trend many more than two data points are required. Google performs slightly more stable but is still behind expectations. The search engines offer no explanation on why the estimates vary. It can be assumed that different versions of the index are used for the estimate.

For activity detection it is required to have at least two observations. While the number of estimated results from one observation can be used to derive a general feeling for the importance of the matter in question we need two observations to get an idea in which direction possible activity is heading. For large numbers Google and Bing provide very volatile results (50 %–70 % change in a matter of hours) which obviously does not reflect the actual state of the Web. The estimated matches proved to be more stable for a smaller result set. Therefore we suggest this method in regards to activity measurement only for topics which yield a small number (a few hundreds to thousands) of results.

As far as trend detection is concerned we would recommend against using this method as the number of results varies too much and a fluctuation might induce a trend which is not actually there though it seems to stabilize if there are enough measurements done. With this method we consequently regard the dimensions time, country and language. We could also add search engine as another dimension. For visualization a world map and a slider for the

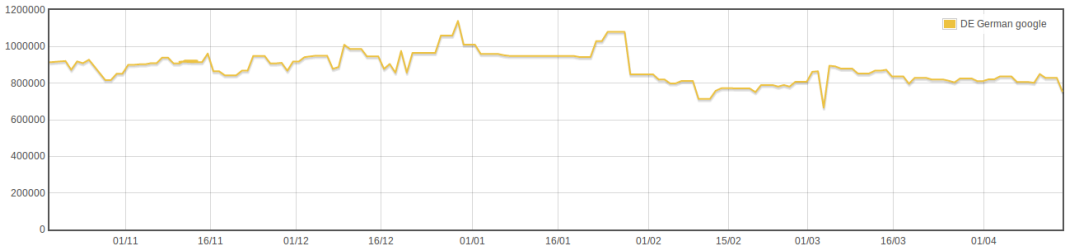

Figure 4 Estimated Matches by Google for "news" in German.

time dimension would be sufficient. The time series can be displayed as graphs. The basis data needed for this method is a byproduct of each query. As almost no processing is taking place this method is easily implemented and has low hardware requirements.

More data is required and further research has to be done to give estimates how much data is necessary for a good trend detection and activity measurement.

3.3 Page Updates

This method requires at least the data of two crawls to work. The method extracts all URLs which are part of the crawls c_{t_0} and c_{t_1}. As the documents were retrieved by the crawler and are stored locally at t_0 and at t_1 we can compare these two versions of the document. If they are dissimilar we assume that an update of either the content or the document structure was made. As the absolute number of updated documents varies as the index of the search engines changes and the ranking of the documents get reevaluated we use the relative update rate of all documents which are part of the intersection of c_{t_0} and c_{t_1}.

Activity is detected easily as an updated document indicates that either by a user, web master or at least by an automated system the effort was undertaken to change content. Counting any change as activity is a generalization we chose on purpose as it drastically improves speed of analysis compared to other approaches and avoids the problem of quantifying the degree "change" of documents. The downside is that things like an automatically generated time stamp on a page would increase the activity rating besides no activity taking place. While this could be fixed with a constant pool of documents where the effects of a rogue document would be canceled out over time we decided to take the risk because we value the relevance rating of the search engines over the occasional misinterpretation of activity.

In regards to trend detection this method is of limited use. A direction can be deducted by the increase or decrease of the activity over time. But currently we can not extract more specific information about the underlying forces of a trend from this method.

First experiments with this method have shown that it provides rather stable results and is opposed to e. g. Estimated Matches not prone to unreliable search engine data. A comparison of two topics showed that "daily news" has a 95 % update rate while the more stable "geoengineering" only has a 25 % update rate aggregated over the whole topic. This method also works independently from the number of documents. The recommended amount of documents for this method is still subject of ongoing research.

The dimensions inspected by this method are time, country, language and search engine. Various aggregations e. g. all languages of a country are possible. Changes of the results by different ways of aggregating are also subject of ongoing research.

Visualization is done by a chart and a world map. The computational complexity depends on the way differences between document versions are detected and treated. A more complex

analysis can be costly. Our approach via a hash allows us to keep the cost in terms of computing time low.

3.4 New Sources

A basic requirement is the results of two crawls c_{t_0} and c_{t_1}. By comparing the URLs provided by the search engines we look for sources which were previously not part of the crawl. The assumption is that if a search engine adds or removes URLs to the result set of a specific query something must have happened to change the relevance rating of this particular or following URLs. Therefore we assume that some kind of activity has happened which results in the changed URLs. We use the search results because big search engines are more capable to analyze a large part of the Web and have a tested infrastructure.

Naturally this method is primarily developed with activity measurement in mind. We can show that a topic like "geoengineering" behaves differently from "news" but due to the way modern search engines handle their index structure this method suffers from the same symptoms as the Estimated Matches method. Search engine provider usually operate with several index structures in parallel. A "current" index is used to answer queries while the next index is built in the background. Adaptations of the current index are done by lists which contain deleted URLs [3]. Currently it is impossible for us to distinguish between actual activity and the switching to a new index at the side of the search engine provider. A high influx of new URLs might suggest a new index structure and in the case of successive large changes activity might be deducted but we can not cancel out the possibility of the reply to the search queries coming from different data centers or index structures. We have to conduct further research to see how this method performs in the long run and how we can detect which index version is delivering the URLs so we can compare changes to the same index.

For many observation this method can be used for trend detection. In the current state the results are too imprecise and fuzzy to detect a direction of the activity.

This method is also working on query basis which allows us to inspect the dimensions time, country, language and search engine. Similar to Page Updates aggregation can be used to get more general information e. g. about the development in a particular country.

For visualization we currently only use a chart as the relationship between available documents and index rebuilding and effect on the result set is not yet clear so a country-by-country comparison does not to seem feasible yet.

4 Conclusions & Outlook

In this paper we presented a new approach to activity monitoring and trend detection in the Web. The new approach consists of a a new concept to acquire data and a number of methods for activity and trend detection. We explained the concept for data acquisition and used an fully implemented prototype to prove its feasibility. Then we continued to show what methods can be applied to the data and which conclusions can be drawn from each method using examples and gathered data.

Activity measurement and trend detection is tightly linked. To develop better methods for trend detection we need the ability to detect activity so that we can research the various causes of the activity. In order to improve our activity measurement methods we need to collect more data and research the relationships of the various dimensions we deal with e. g. search engine.

We plan to use the developed methods to analyze the performance of various search engines in regards to index variance and refresh rate. Also we need to evaluate the stability of the results of the search engine providers so that we can give a qualified suggestion which providers are best fit for our methods.

In regards to trend detection we research the average live span of trends in various topics. Also a field of our research is the clustering of trend terms via correlation analysis and spatial analysis based on the downloaded documents.

Further technologies which we would like to include in future work are positional indexes. Positional indexes enables the detection of terms which consist of more than one token (e. g. New York) by searching for combinations of terms which appear close to each other (n-grams). The proposed text-based methods for activity and trend detection are unaffected by a positional index as an abstraction layer can be built to keep the input format unchanged.

We plan to validate the methods using the Reuters Information Retrieval Text Research Collections though there is no such thing as a gold standard for trend detection on unstructured data.

References

1 Wangberg, Silje C., et al. *Relations between Internet use, socio-economic status (SES), social support and subjective health.* Health promotion international 23.1 (2008): 70–77.

2 Miniwatts Marketing Group. *Internet World Stats.* 31 May 2011, accessed 16 April 2014

3 Manning, Christopher D. and Raghavan, Prabhakar and Schütze, Hinrich. *Introduction to information retrieval.* Cambridge university press Cambridge, USA, 2008

4 Sun, Yang and Zhuang, Ziming and Giles, C. Lee. *A large-scale study of robots. txt.* Proceedings of the 16th international conference on World Wide Web. ACM, 2007.

5 Papineni, Kishore, et al. *BLEU: a method for automatic evaluation of machine translation.* Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.

6 Fung, Pascale. Extracting key terms from Chinese and Japanese texts. Computer Processing of Oriental Languages 12.1 (1998): 99–121.

7 McCown, Frank and Nelson, Michael L. *Search engines and their public interfaces: which apis are the most synchronized?* Proceedings of the 16th international conference on World Wide Web. ACM, 2007