Crosslanguage Blog Mining and Trend Visualisation

Andreas Juffinger
Know-Center
Inffeldgasse 21A/II
Graz, Austria
ajuffinger@know-center.at

Elisabeth Lex
Know-Center
Inffeldgasse 21A/II
Graz, Austria
elex@know-center.at

ABSTRACT

People use weblogs to express thoughts, present ideas and share knowledge, therefore weblogs are extraordinarily valuable resources, amongs others, for trend analysis. Trends are derived from the chronological sequence of blog post count per topic. The comparison with a reference corpus allows qualitative statements over identified trends. We propose a crosslanguage blog mining and trend visualisation system to analyse blogs across languages and topics. The trend visualisation facilitates the identification of trends and the comparison with the reference news article corpus. To prove the correctness of our system we computed the correlation between trends in blogs and news articles for a subset of blogs and topics. The evaluation corroborated our hypothesis of a high correlation coefficient for these subsets and therefore the correctness of our system for different languages and topics is proven.

Categories and Subject Descriptors

H.4.m [**Information Systems**]: Web Content, Blogosphere

General Terms

Algorithms, Experimentation

1. INTRODUCTION

Weblogs are used to express thoughts, to present ideas and to share knowledge, so topics posted in weblogs are highly applicable for trend analysis. In our project for the Austria Press Agency (APA)[1] we aim to exploit the "collective wisdom" on the web in a sense of comparing the "collective opinion" against the "news opinion". The blogosphere[1] presents an opportunity for us to understand the influence of certain news to the public and their propagation by analysing unsolicited feedback[3] in blogs. The ability to derive the propagation of news messages around the world makes it necessary to support different languages.

Related work in the field of blog mining, influence in blogs and propagation, and blog visualisation reveals a strong evidence that the blogosphere correlates with the real world based on quantitative and qualitative analysis. Drezner and Farrel were able to show[2] that there is an interdependency between blogs and the real world. BlogPulse[2], a blog analysis service, use the percentage of all posts concerned with a topic of interest to show trends in blogs. Google Trends[3], a visualisation of search patterns, depicts the number of times a topic occurs in Google News as a reference line.

2. CROSSLANGUAGE BLOG MINING

Blog mining is defined as the integrated discipline of web and text mining, with refined techniques from social network analysis for the specific structure and content of weblogs. Our blog mining system implements a high performance web miner and can incrementally load and parse blog sites. During the blog parsing step we transform the unstructured content into a structured blog entry and thus prepare the content for our text mining unit. For this purpose we developed a semiautomatic blog site parser based on relative XPath queries[6]. From the structured blog entries we extract named entities and the language of the blog posts. Title and content of the posts are then indexed in language specific indices with Apache Lucene[4].

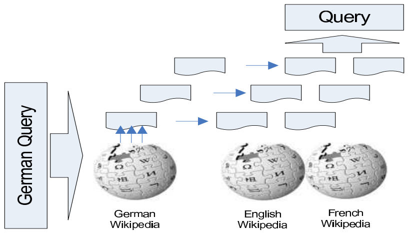

Figure 1: Query Translation with Wikipedia

The applied methodology for crosslanguage retrieval is based on wikipedia statistics as outlined in [4]. The applied workflow for query translation for the blog mining system is shown in Figure 1. Our primary data source is the German APA news corpus, so the query language is restricted to German. Each query is used to search in the German Wikipedia index. From the top fifty result documents we extract the linked Wikipedia articles for each target language (English, French, Spanish and Italian). The result of this extraction step is a set of relevant articles in every language. These articles are then used to extract the statistically most significant terms for each language in the context of the query. The significant terms are then used to search the language specific blog index.

[1] http://www.apa.at

[2] http://blogpulse.com

[3] http://trends.google.com

[4] http://lucene.apache.org

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

3. BLOG TREND VISUALISATION

The Blog Trend Visualisation is embedded in the APA Labs[5] framework. APA Labs enables to search the news repository, to navigate results through result lists and to visually analyse search query results and article content[5]. The Blog Trend Visualisation shows APA articles and blog entries over a specific time period. Besides, the articles are also available in a tabular view. The Blog Trend Visualisation shows trends and language specific propagation over time at a glance and serves as a starting point for qualitative analyses. Figure 2 gives an example of the visualisation for the search query term "Bush". The time period is limited to the last 60 days. Three days are then summarised and depicted by a coloured bar. Symbols for blog entries and articles are shown on the bar if results are present for this timeslot.

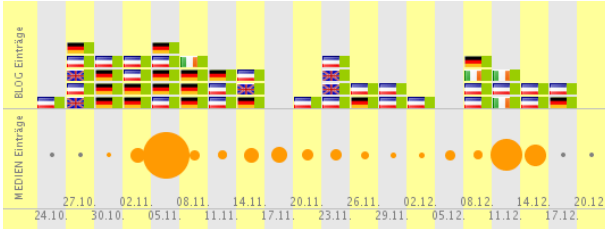

Figure 2: Visualisation for query "Bush"

The blog entries are shown in the top section of the visualisation. Each blog entry is represented by an icon consisting of a green rectangle and a flag that denotes the language of the blog entry. In Figure 2 German, English, French and Italian blog entries are present. Each blog entry icon provides a tooltip that shows the title. The icon can also be used to navigate to the associated blog entry on the external blog site.

Orange circles represent the news articles in the bottom section of the visualisation. The size of a circle corresponds to the number of articles found to the search query for the particular timeslot. Clicking a circle restricts the articles in the tabular view to articles of the selected timeslot.

4. EVALUATION

To evaluate our crosslanguage system we analysed the correlation between news articles and a selection of blogs. For a high correlation coefficient between news articles and blog entries the selection of blogs is crucial. Consequently, an automatic approach for blog selection was not an option. For the evaluation we handselected about 40 blogs by popularity, actuality and significance to guarantee that the blogs are active and current. Furthermore, only blogs dealing with current events were selected to guarantee a high correlation with news articles. The blogs are equally distributed over topics and languages because experiments also revealed that one single blog can glut the system and overlap the originally correlated functions. We evaluated the correlation for 15 person names, 15 location names, and 15 arbitrary query terms. Figure 3 shows the mean, standard deviation, minimum and maximum of the correlation between the number of blog entries and number of news articles.

The evaluation for different languages (top image) revealed that the system performs equally good for German, French and English, although the correlation for English was lower caused by the selection of blogs. For the multilanguage experiment we achieved a smaller standard deviation due to the higher number of blog entries in the input data (all blog entries vs. blog entries in a specific language) what led to a higher statistical robustness. The overall correlation coefficient for multilanguage retrieval was 0.55 (bottom image, rightmost bars) with a standard deviation of 0.18, a maximum of 0.93 and minimum of 0.02. Consequently, the system is robust for the different types of query terms and languages.

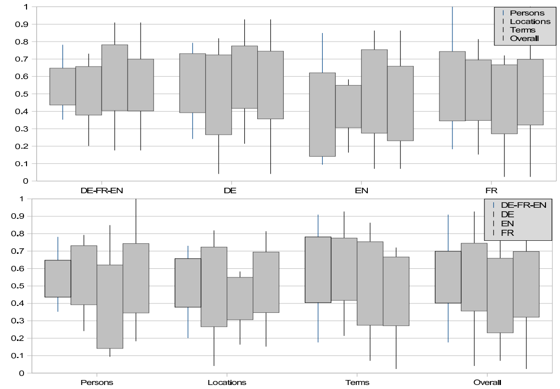

Figure 3: Correlation between News and Blogs

5. CONCLUSIONS

Our crosslanguage blog mining and trend visualisation system enables to identify trends in blogs in analogy to a news article repository. The visualisation shows the diffusion of a topic of interest and serves as a starting point for in-depth qualitative analyses. Additionally, we are able to show a strong correlation between news articles and selected blogs across different languages. Consequently, we can continuously validate our system and identify topic drifts even in foreign languages.

6. ACKNOWLEDGEMENTS

The Know-Center is funded within the Austrian COMET Program under the auspices of the Austrian Ministry of VIT, Austrian Ministry of WA and by the State of Styria.

7. REFERENCES

[1] N. Agarwal and H. Liu. Blogosphere: Research issues, tools, and applications. *SIGKDD Explorations*, 2008.

[2] D. Drezner and H. Farrell. The power and politics of blogs. In *Proc. of the APSA Conf.*, 2004.

[3] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo. Deriving marketing intelligence from online discussion. In *Proc. of the Knowldege Discovery and Data Mining Conf.*, 2005.

[4] A. Juffinger, R. Kern, and M. Granitzer. Crosslanguage retrieval based on wikipedia statistics. In *Proc. of CLEF 2008 Workshop, Aarhus*, 2008.

[5] W. Kienreich, E. Lex, and C. Seifert. APA Labs: an experimental web-based platform for the retrieval and analysis of news articles. *Proc. of ICADIWT*, 2008.

[6] M. Kowalkiewicz, M. E. Orlowska, T. Kaczmarek, and W. Abramowicz. Robust web content extraction. In *Proc. of the 15th int. conf. on World Wide Web*, 2006.

[5] `http://www.apa.at/labs`