Blog Credibility Ranking by Exploiting Verified Content

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

Michael Granitzer
Know-Center, Graz
Inffeldgasse 21a
Graz, Austria
mgranitzer@know-center.at

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

ABSTRACT

People use weblogs to express thoughts, present ideas and share knowledge. However, weblogs can also be misused to influence and manipulate the readers. Therefore the credibility of a blog has to be validated before the available information is used for analysis. The credibility of a blogentry is derived from the content, the credibility of the author or blog itself, respectively, and the external references or trackbacks. In this work we introduce an additional dimension to assess the credibility, namely the quantity structure. For our blog analysis system we derive the credibility therefore from two dimensions. Firstly, the quantity structure of a set of blogs and a reference corpus is compared and secondly, we analyse each separate blog content and examine the similarity with a verified news corpus. From the content similarity values we derive a ranking function. Our evaluation showed that one can sort out incredible blogs by quantity structure without deeper analysis. Besides, the content based ranking function sorts the blogs by credibility with high accuracy. Our blog analysis system is therefore capable of providing credibility levels per blog.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation

Keywords

Blog, Credibility, Web 2.0

1. INTRODUCTION

A Web site has to be credible in order to ensure that the provided information and services are credible. Due to insufficient Web content quality control mechanisms, information on Web sites is often not correct or even misleading.

As showcase one can use the announcement of Jimmy Wales, the founder of Wikipedia, to introduce editors in Wikipedia to check the credibility of content before articles are published. Especially the credibility of user generated content like blogs, news groups etc. is questionable. Credibility is defined in [8] as the ability to inspire belief or trust and as information accuracy and veracity[12]. According to the Stanford Guidelines for Web Credibility[1] one should make it easy to verify the accuracy of the information on Web sites. User studies showed[4] that users judge a Web site credible primarily based on structural and author specific elements. Note that in contrast to standard Web sites, people can be less controlled and are even harder to trace in the blogosphere[1] and other Web2.0 applications. A blogger, for example, is only identified by the username, often nicknames, with no meaning. Because all blogs share a similar structure - the most important are title, date, and content - structural information is less significant for credibility analysis.

Most work in the field of blog credibility has focused on trust and reputation networks [6, 2, 17] motivated by Web site authority ranking. Kleinberg [11] and Page et al. [16] developed a ranking mechanism for Web search engines based on the idea of citation networks [22]. Citations in scientific papers are clearly a measure of credibility because published papers are reviewed and therefore no artificial papers exist to boost the citation count. But citation networks on the Web can be spammed and consequently loose their credibility if not validated. A popular example for this technique of spamming a citation network on the Web are linkfarms. Linkfarms are an accumulation of Web sites or domains and are used to link as many hyperlinks as possible to another Web site. Linkfarms are used to manipulate search engines ranking techniques. A popular example of this is the Google PageRank algorithm [14]. The PageRank algorithm ranks the relevance of a Web site by the quality and quantity of the sites that link to it. In PageRank each Web page has a measure of authority, the authority of a page is proportional to the sum of the authority scores of pages linking to it. Early versions of the PageRank algorithm have been sensible to link spamming what has been exploited. Nowadays, Google uses about 200 signals to rank search results[2]. Of course content based source are also spammable but link spamming is harder to detect as it can boost the ranking of the target pages without changing the content. Because each source for credibility assessment can be influenced by

---

[1]http://credibility.stanford.edu/guidelines
[2]http://googleblog.blogspot.com/2008/03/why-data-matters.html

a spammer, we therefore recommend especially for Web 2.0 applications and blogs to always take at least two sources into account for credibility assessment. For a spammer it is disproportionate harder to influence two or more parameters.

For credibility assessment in blogs one has also to consider a special type of spamming, namely Splogs (Spam Blogs). These blogs are artificially created and used to promote associated sites to boost the search engine rankings of those target sites. The content of Splogs is often automatically copied from popular news portals and enriched with links to the target site. This technique improves thereby the authority of the page. Splogs are very hard to distinguish from credible blogs and news portals.

The blogosphere presents an opportunity for us to understand the influence of certain news to the public and their propagation by analysing unsolicited feedback[5] in blogs. In our project for the Austrian Press Agency(APA)[3] we developed a system to analyse and visualise blogs over different languages. The aim of the blog analysis system is to illustrate the language specific trends of a topic discussed in blogs over time. For deeper analysis of these trends it is necessary to distinguish between highly credible and gossip blogs. But how can we distinguish between gossip and credible information in our blogs? Respectively, can we provide an automatic credibility level to the user by analysis the information and content in blogs?

Unfortunately we are limited to credibility assessment from content and temporal distribution only, due to the data nature of our project. As discussed earlier structural information as in [21] cannot be used because our blog selection strategy sorts out clearly incredible blogs like "advertise blogs" before we start the credibility analysis. Author information is currently not available. Also, network analysis based on trackbacks is meaningless due to the small number of links inbetween our blogs. On the other hand the amount of blogs processed for APA is too big for manual credibility assessment. However we benefit from the fact that we can access a validated reference news corpus. For our blog analysis system we derive the credibility from two dimensions: firstly, the quantity structure of blog entries, and secondly, the content of the entries. We compare statistical properties of the quantity structure and blog content with the appropriate properties of a reference news corpus. Based on these properties we then rank the blogs with three different credibility values: "1" as "highly credible", "2" as "average credible" or "unspecified" and "3" as "little credible". In this context we refer to "little credible" as gossip.

The rest of this paper is structured as follows: In Section 2 we introduce the overall system in which the credibility assessment is embedded. Then Section 2.3 introduces our credibility level extraction process and methodology. The building blocks and their evaluations are then outlined in Section 3 and 4, before we conclude this paper in Section 5.

2. BLOG ANALYSIS SYSTEM

The blog analysis system consists of three main parts, a blog crawling and mining unit, a blog trend visualisation and the proposed blog credibility ranking process. Within the blog analysis system it is possible to formulate search queries and visually analyse the result sets of blog entries and news

articles. With the credibility ranking process we provide additional information to identify the credibility level. The particular components of the blog analysis system are described in more detail in the next sections.

2.1 Blog Crawling and Mining

Our blog crawling application is based on a high performance web miner [10]. The current blog crawler extends the Web miner with incremental blog load and parse functionality. Our blog parser uses relative XPath queries as outlined in [13]. The parser works top down. Firstly, the section with blog entries is extracted and the content of irrelevant sections, like blog archive or blog navigation items, is cut off. Secondly, the separate blog entries are extracted from the content section. Finally, the relevant information of each remaining single blog entry is extracted. Therefore we transform the unstructured content into a structured blog entry and prepare the content for text mining. The extracted features are title, date, author, content, language, tags and permanent link. The blog entries are then indexed with Apache Lucene[4], a high-performance, full-featured text search engine library.
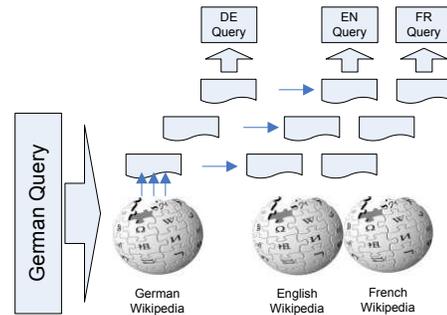

Figure 1: Query translation with Wikipedia.

The blog mining unit is capable of retrieving blog entries of different languages according to a German query. The applied methodology for crosslanguage retrieval is based on Wikipedia statistics as outlined in [9]. The workflow for query translation used in 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. The term weights are computed by methods from pseudo relevance feedback and query reconstruction. Based on these weights we take the top terms whose weights are at least 85% of the weight of the most relevant term. Usually this leads to 5-10 query terms for each language specific blog index. Because person names do not have to be translated they are directly used as query terms in each blog index.

2.2 Blog Trend Visualisation

The blog analysis visualisation is part of the APA Labs[5] framework. Within the APA Labs it is possible to search the news article repository, to navigate search result lists and to visually analyse search query results and article content [15]. The blog analysis module extends the APA Labs framework with a search in international blogs. The blog analysis visualisation depicts news articles and blog entries over a defined time period on a time axis. Clicking an icon in the visualisation either directly opens the according blog entry in a separate browser window or fetches the corresponding news articles from the news repository. The aim of the blog analysis visualisation is to illustrate the language specific distribution of topics discussed in blogs over time. The German news articles serve as a baseline to compare the quantity structure of a topic in the news and the blog world. In Figure 2 an example of the blog analysis visualisation is given. Because the visualisation points out the diffusion of a topic of interest over time it serves as a starting point for in-depth qualitative analyses.
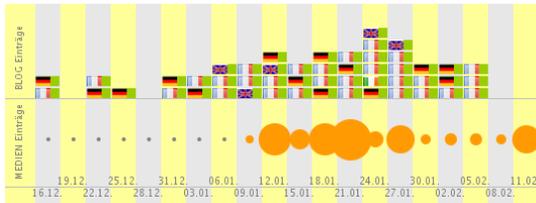

Figure 2: Example of the blog analysis visualisation.

The blog analysis visualisation will be extended with the credibility ranking information extracted according to the process as outlined in Section 2.3. The current mock up, as shown in Figure 3, denotes "1" as "highly credible", "2" as "average credible" or "unspecified" and "3" as "little credible" or "gossip" with a small number next to each blog entry.
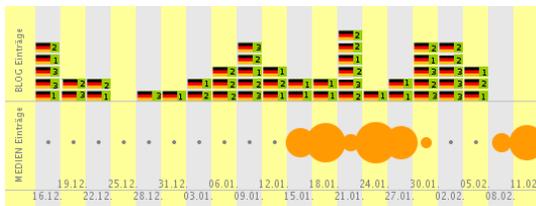

Figure 3: Example of blog visualisation with blog credibility level.

The credibility information will support the analysts to focus on information from a certain credibility level. This enables the analysis of trends on a certain credibility or objectivity level. Besides, a comparison of trends over different levels is possible. Such a comparison is highly applicable for media resonance analysis. For example, a company is interested in the resonance to their new product. For this it is highly valuable to get feedback from objective and credible resources like blogs from experts in the field. On the other hand it might be also interesting how the product is discussed in gossip blogs to work against baseless rumours.

2.3 Credibility Ranking Process

The blog credibility ranking procedure starts with a search query in the verified APA news article repository and in the blog repository. We use the query results from the news corpus as a basis for computation and alignment of the blog entries for credibility ranking. Unfortunately the news articles are only available in German and therefore our credibility ranking process is currently restricted to German. The process of blog credibility ranking is shown in Figure 4.
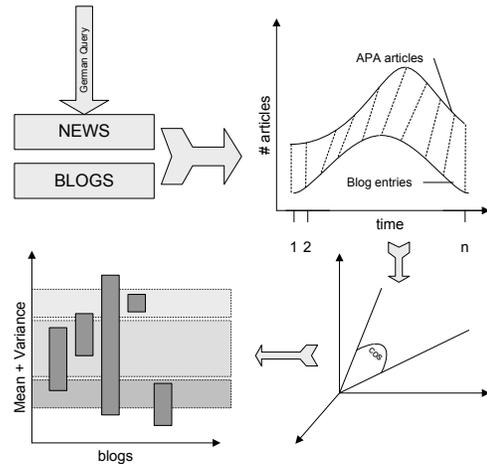

Figure 4: General process of blog credibility ranking.

A German query serves as input to search in the APA news repository and in the blog index. Both modules output documents in German language. These documents are then subject to a two-stage process. Firstly, a quantity structure analysis and filtering step based on a timewarping procedure and correlation coefficient calculation is performed (see Section3). Secondly, a content similarity evaluation provides a credibility ranking (see Section 4).

3. TIME WARPED CORRELATION

Observations of the temporal distribution of news articles and blog entries over time, as shown in Figure 5 and 6, revealed a correlation between these two medias. On closer examination, the time series of blog entries and news articles to a search query revealed a strong correlation only when the quality of the blogs is high. The quality is high when the blogs are not spammed with advertising or meaningless blog posts.
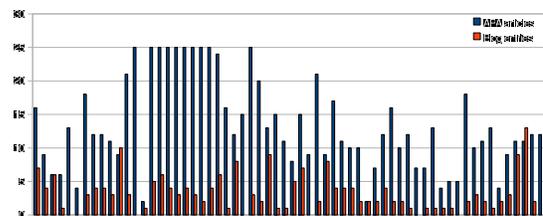

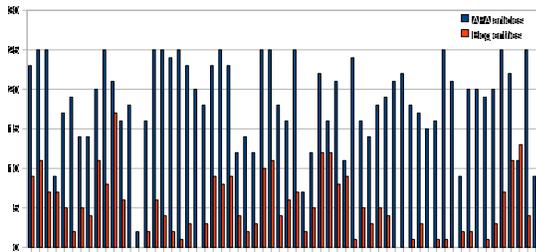**Figure 5: Temporal distribution of the news articles and the blog entries for search query "Obama".**

53

Figure 6: Temporal distribution of the news articles and the blog entries for search query "Frankreich".

Figure 7: Quantity structure of different queries.

structure of the blog set and the news corpus.

Figure 8: Correlation calculation for different queries.

One major problem of aligning different corpora over time is that they usually differ in the amount of items per day. Also, they are sometimes shifted in time, stretched or compressed. This is shown in Figure 4. For optimal alignment of the different time series we utilise the dynamic time warping (DTW) algorithm. This algorithm is often used to determine time series similarity, classification, and to find corresponding regions between two time series. The standard DTW algorithm has a quadratic time and space complexity. For performance reason we implemented a FastDTW based on the ideas of Fu et al. [3] and an exploitation limit by the Sakoe-Chiba Band[19] and therefore the algorithm performs nearly in linear time. From the aligned time series we then compute the correlation coefficient.

In the context of credibility ranking, this correlation test filters out blogs with significantly different quantity structure for a certain query in comparance to the validation corpus. With a "leave one out" technique we compare the correlation coefficient of the news distribution with the blog set of interest. Whenever we achieve a significantly higher coefficient without a blog, this blog is sorted out. The robustness of the correlation coefficient against a constant offset is thereby helpful, because possibly credible blogs with a constant amount of entries per day do not have an impact on the correlation coefficient and are therefore not sorted out. Blogs with actual events shortly after or before the news papers remain also, due to the time warping. We only sort blogs out with a negative influence on the correlation and that are the blogs with a completely different distribution over time. Figure 7 shows the number of blog entries and news articles according to 30 different queries in a specific time period. One can clearly see that there is a very strong correlation of the quantity structure.

However this process does not filter out blogs with wrong content at the right time. Therefore another strategy has to be applied which takes the actual content of blog posts into account. Our solution for this problem is proposed in the Section 4.

3.1 Evaluation

To ensure a similar performance over different queries we evaluated the performance of our correlation calculation for various named entity types. For this 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 used to guarantee a high correlation with news articles. Besides, the blogs are equally distributed over topics and languages. This selection guarantees a very high correlation between the quantity
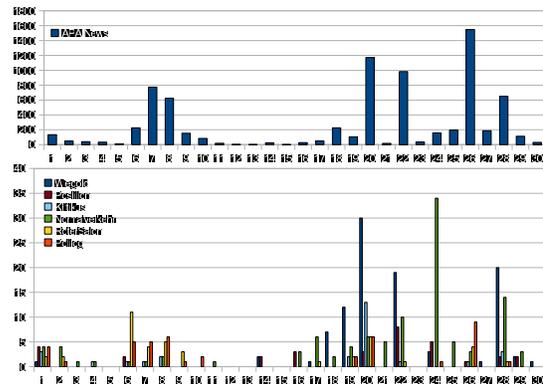
We evaluated the correlation for 15 person names, 15 location names, and 15 arbitrary query terms. Figure 8 shows the mean, standard deviation, minimum and maximum of the correlation between the number of blog entries and number of news articles. The evaluation shows that our strategy reveals the expected high correlation for this dataset. Also, the correlation is quite similar for all types of named entities. Therefore we can use any kind of queries for the correlation test.

4. CREDIBILITY RANKING

After the correlation filter we end up with a set of blogs with appropriate quantity structure. This set of blogs is the basis for deeper content based analysis and credibility ranking. We aim to assign one credibility level per blog based on content analysis. We analyse the content of the blogs and measure the similarity with the content of the news articles. For analysis we take all news articles into consideration that result from a search query - all articles are relevant. The centroid of the articles in the Vector Space Model[20] is then taken as a representative for the news corpus. Regarding the blog entries we have to follow a different strategy because it is not mandatory that all blog posts in a particular blog relate to the search query. However we want to rank the whole blog. Therefore we developed and evaluated two strategies to deal with this problem:

S1 Only the blog entries relevant to the search query are used to rank the blog.

S2 All blog entries in the time period are used to rank the blog.

4.1 Natural Language Processing

From our experience, nouns generally cover the thematic information and verbs the association within the topic. Therefore part of speech information is needed. To vectorise the news articles and blog entries we perform a natural language processing step with lemmatisation and part of speech tagging. For this, we use an information extraction module developed at our institute. The module is based on the Java package OpenNLP[6] and performs tokenisation, sentence splitting and part-of-speech (POS) tagging. The sentence splitter is thereby based on a maximum entropy model and the POS tagger utilises a dictionary of tags and a trained model to add the POS tags to each token in the particular sentence. The output of the tagger follows a style proposed by the so-called "Penn Treebank Project"[7]. After the tokenisation and the POS tagging we apply stemming to reduce the various morphological variants of the words to their common word stem. Our stemming application is based on Snowball[8], a language that defines stemmers and enables to generate fast stemmer programs in Java.

4.2 Centroid Cosine Similarity

The result of the NLP Module is a list of stemmed terms per document with the absolute term frequency (ATF) of each term. In the vector-space model the documents are regarded as vectors in the document term-space. The document term vectors are then TF-IDF weighted and normalised to unit length. We calculate the centroid vector of all news articles and the centroid vector per blog. The computation of the centroid vector is shown in Equation 1 whereas S denotes the number of documents and d represents the document term vector[7].

$$C = \frac{1}{|S|} \sum_{d \in S} d \qquad (1)$$

The centroids are then again normalised and at last we compute the cosine similarity inbetween the blog centroids and the news centroid. The cosine similarity is a common content based similarity measure [18]. The mathematical definition of the cosine similarity is given in Equation 2 whereas d denotes the vector space representation of the documents [7].

$$similarity = cos(d_i, d_j) = \frac{d_i.d_j}{\|d_i\|_2\|d_j\|_2} \qquad (2)$$

For each query we compute one similarity value between the news corpus and each blog. Then we calculate the mean and variance of these similarity values over a predefined set of queries. These queries all belong to the same topic for which the credibility ranking should be calculated. For example to calculate the credibility for the topic politics, the queries contain only politician names, political regions, and political concepts. For the similarity calculation we take different part of speech elements into account:

- Nouns: nouns generally cover the thematic information
- Verbs and adjectives: both denote the association within the topic

4.3 Credibility Assignment

The similarity values are used to assign one out of three credibility levels to each blog. In the first step we take only the similarity values computed based on nouns into account. All mean and variances are then compared with each other and we use a threshold to sort out those blogs with the lowest similarity values. The threshold is derived from the interval defined in Equation 3.

$$t = \left[0, min(sim_j) + \frac{1}{2} * (max(sim_i) - min(sim_j)) \right] \qquad (3)$$

All blogs below the threshold are assigned to the credibility level 3 ("little credible"). The same procedure is applied to the similarity values computed based on verbs and adjectives. Experiments showed that the low credibility states that these blogs are either dealing with a different topic (from nouns) or are in a completely different association with the topic (from verbs and adjectives).

From the remaining blogs we compare the similarity values computed without restrictions on part of speech types. All blogs within the interval given in Equation 4, are assigned to the credibility level 1 ("highly credible"), if their variance is less than 0.05.

$$t = \left[min(sim_j) + \frac{1}{2} * (max(sim_i) - min(sim_j)), 1 \right] \qquad (4)$$

All blogs which neither fall into level 1 nor 3 are finally assigned with level 2 ("average credible" or "unspecified").

4.4 Evaluation

To evaluate our system we perform a selection of queries on a sufficient big amount of blogs and compute the credibility ranking according to the earlier described process and different scenarios defined in S1 and S2.

The evaluation revealed that we cannot retrieve a credibility ranking from the similarity values between the news corpus and only the relevant blog entries (Scenario S1). For example, if one blog entry perfectly matches the news result set, one cannot state anything about the full blog. Therefore we use only Scenario S2, all blog entries per time period, for our credibility assignment strategy.

Figure 9 shows the cosine similarity mean and variance for relevant blog entries.
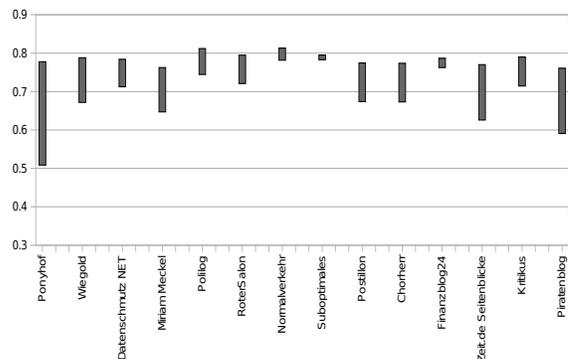

Figure 9: Cosine similarity mean and variance for relevant blog entries.

Figure 10 shows the similarity values for nouns only. The blogs below the threshold, depicted as highlighted area in the figure, can clearly be identified ("Ponyhof", "Wiegold", "Postillion", "Chorherr", "Finanzblog", "Seitenblicke", "Kritikus" and "Piratenblog"). It is noticeably that all blogs, except "Ponyhof", fall clearly below the threshold. The small variance shows that the blogs consistently deal with the topic in a different way.
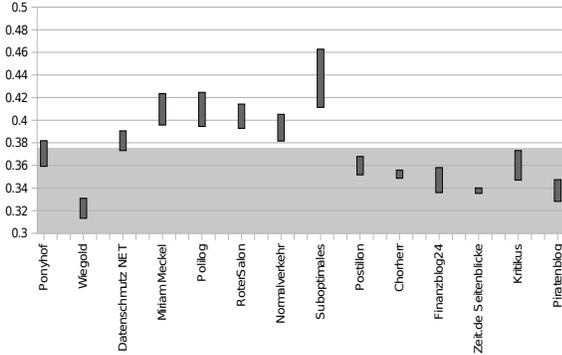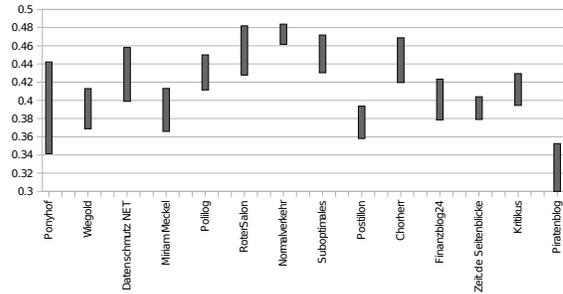

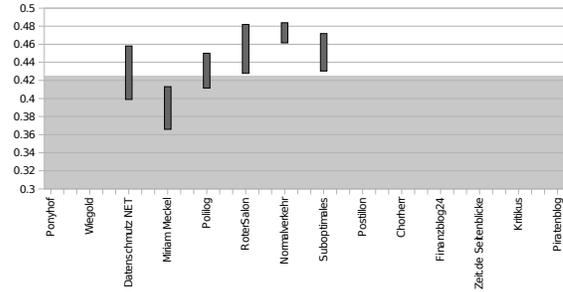**Figure 10: Centroid cosine similarity mean and variance for nouns.**

Figure 11 depicts the similarity values for verbs and adjectives only. As a comparison in the Figure11(a) all blogs are shown (without noun filter). The calculation of the threshold based on this blog set would lead to nearly no additional assignments. This is due to the correlation between nouns, verbs and adjectives in natural language texts. Naturally, documents off topic have a small similarity for both term types. Therefore, in this step we take only unlabelled blogs into account. Figure11(b) shows the similarity values only for the remaining six blogs (with noun filter). On this set we again apply the threshold and are able to assign the "little credible" level to the "Miriam Meckel" blog.

Figure 12 shows the similarity values for setting S2. This uses all blog entries in the period without part of speech restrictions. In Figure12(a) the mean and variance of centroid cosine similarity values for all blogs are shown. It can clearly be seen that without our preselection based on nouns and verbs no assignment of a specific credibility level to each blog is possible. However, with preselection, the remaining blogs are divided into "highly credible", "medium credible" or "unspecified" using the threshold given in Equation4. The blogs above the threshold (see Figure12(b)) and with low variance are classified as highly credible ("Polilog", "Roter Salon", "Normalverkehr", and "Suboptimales"). According to our process the remaining blogs fall into the middle. Consequently we assign them either being unspecified or average credible.

To verify the results domain experts also used our credibility levels to rank the blogs. All nine blogs assigned with "little credible" by our approach were also labelled as "little credible" by the experts. Unfortunately, two "little credible" blogs were not correctly assigned by our system ("Polilog" and "Roter Salon"). We closer investigated the two false ranked blogs and figured out that their structure as well as the wording is quite similar to the news articles. Thats why they were falsely assigned even though their content is not credible at all. To assign a correct credibility value
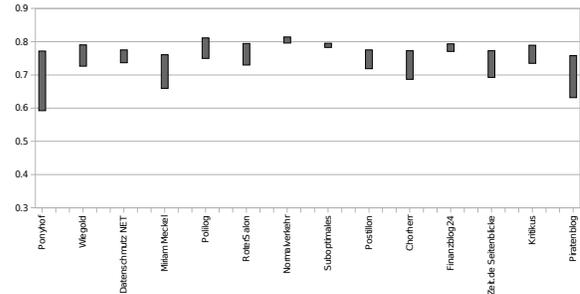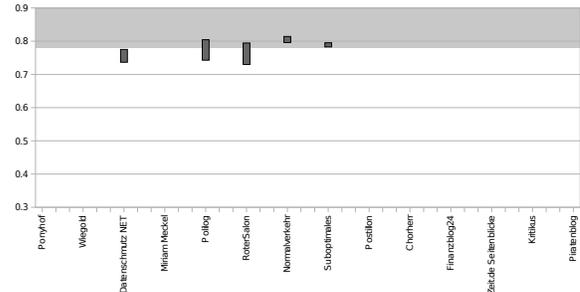

(a) Centroid cosine similarity.

(b) Similarity threshold.

Figure 11: Cosine similarity mean and variance for verbs and adjectives.

(a) Centroid cosine similarity.

(b) Similarity threshold.

Figure 12: Cosine similarity mean and variance for all term types.

to these remaining blogs it is necessary to apply more complex, semantic content analysis methods. Currently we work on a strategy to exploit verb-noun tuples to derive a five

level ranking and to improve the accuracy. Another strategy would be to use subject-predicate-object triples to consider even semantic relations. We assume that such an integrated approach would make a preselection based on nouns and verbs redundant.

To sum up, our system assigned 12 out of 14 blogs correctly to the appropriate credibility level. The precision for level 1 is therefore 0.5, for level 2 it is 1.0 and for level 3 it is also 1.0, leading to an average precision of 0.83. This relatively high precision is yet not very significant due to the small number of considered blogs.

5. CONCLUSION

In conclusion, our blog credibility ranking system enables to automatically rank blogs by three levels of credibility. We estimate the blog credibility by exploiting the quantity structure and the content similarity in reference to a German news corpus. Adding the credibility information to our existing blog analysis system will support analysts and users to focus on blogs of a specific credibility level. The evaluation results indicate a high quality assignment to the three credibility levels with an average precision of 0.83 on 14 blogs, yet more exhaustive evaluations are needed to allow statements about the generalisation capabilities of our approach.

6. ACKNOWLEDGEMENTS

7. REFERENCES

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

[2] A.Kale, A. Karandikar, P. Kolari, A. Java, T. Finin, and A. Joshi. Modeling trust and influence in the blogoshpere using link polarity. In *Proc of Int. Conf. on Weblogs and Social Media*, 2007.

[3] A. W. chee Fu, E. Keogh, L. Y. H. Lau, and C. A. Ratanamahatana. Scaling and time warping in time series querying. *VLDB*, 2005.

[4] B. Fogg, J. Marshall, and T. K. et al. Web credibility research: a method for online experiments and early study results. In *CHI '01: CHI '01 extended abstracts on Human factors in computing systems*, pages 295–296. ACM, 2001.

[5] 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.

[6] D. Gruhl, D. Liben-Nowell, R. Guha, and A. Tomkins. Propagation of trust and distrust. In *Proc. of the 13th international conference on World Wide Web*, 2004.

[7] E.-H. Han and G. Karypis. Centroid-based document classification: Analysis and experimental results. In *PKDD '00: Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, 2000.

[8] M. Iding, B. Auernheimer, and M. E. Crosby. A metacognitive approach to credibility determination. In *Proc. of the 2nd Workshop on Information Credibility on the Web*, 2008.

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

[10] A. Juffinger, T. H. Neidhart, M. Granitzer, R. Kern, A. Weichselbraun, G. Wohlgenannt, and A. Scharl. Distributed web2.0 crawling for ontology evolution. - in: International journal of internet technology and secured transactions. *International Journal of Internet Technology and Secured Transactions*, 2008.

[11] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46, 1999.

[12] E. B. Klemm, M. Iding, and T. Speitel. Do scientists and teachers agree on the credibility of media information sources? *International Journal of Instructional Media*, 28, 2001.

[13] 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.

[14] A. N. Langville and C. D. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, 2006.

[15] E. Lex, C. Seifert, W. Kienreich, and M. Granitzer. A generic framework for visualizing the news article domain and its application to real-world data. *Journal of Digital Information Management*, pages 434–442, 2008.

[16] L. Page, S. Brin, R. Motwani, and T.Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[17] M. Pujol, R. Sangesa, and J. Delgado. Extracting reputation in multi agent systems by means of social network topology. In *Proc. of the first international joint conference on Autonomous agents and multiagent systems*, 2002.

[18] L.-Q. Qiu and B. Pang. Analysis of automated evaluation for multi-document summarization using content-based similarity. In *ICDS '08: Proceedings of the Second International Conference on Digital Society*, pages 60–63, 2008.

[19] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1978.

[20] G. Salton. *Automatic text processing: the transformation, analysis, and retrieval of information by computer.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[21] N. Wanas, M. El-Saban, H. Ashour, and W. Ammar. Automatic scoring of online discussion posts. In *WICOW '08: Proceeding of the 2nd ACM workshop on Information credibility on the web*, pages 19–26, New York, NY, USA, 2008. ACM.

[22] E. Garfield, I. Sher, and R. Torpie. *The Use of Citation Data in Writing the History of Science* , Institute for Scientific Information Inc., Philadelphia, Pennsylvania, USA, 1984.