2013 IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT)

Identifying Domain Experts in the Blogosphere - Ranking Blogs based on Topic Consistency

Philipp Berger
Hasso-Plattner-Institute
Potsdam, Germany
philipp.berger@hpi.uni-potsdam.de

Patrick Hennig
Hasso-Plattner-Institute
Potsdam, Germany
patrick.hennig@hpi.uni-potsdam.de

Christoph Meinel
Hasso-Plattner-Institute
Potsdam, Germany
office-meinel@hpi.uni-potsdam.de

***Abstract*—Current ranking algorithms, such as PageRank, Technorati authority, and BI-Impact, favor blogs that report on a diversity of topics since those attract a large audience and thus more visitors, links, and comments. On the other side, niche blogs with a very specific topic only attract a small audience and thus have only a small reach. This results in a low ranking from today's blog retrieval systems.**

We argue that the consistency of a blog, i.e. how focused an author reports on a single topic, is a sign for expert knowledge. To find these blogs is particular important for other domain experts to identify blogs that they would like to follow and stay in active contact. To ease the retrieval of expert blogs, i.e. to separate them from the mass of blogs that report on random topics, we introduce a metric for blogs based on topic consistency. We divide the consistency ranking in four different aspects: (1) intra-post, (2) inter-post, (3) intra-blog, and (4) inter-blog consistency.

By evaluating the metric with a test data set of 12,000 crawled blogs, we demonstrate the plausibility of our approach.

I. Introduction

Weblogs, called *blogs*, are one of the most popular "social media tools" of the World Wide Web (WWW) [7]. They are specialized, but easy-to-use, content management systems. Blogs focus on frequently updated content, social interactions, and interoperability with other Web-authoring systems.

The actual power of blogs evolves through their common superstructure, i.e. a blog integrates itself into a huge think tank of millions of interconnected weblogs, called blogosphere that creates an enormous and ever-changing archive of open source intelligence [20].

Through the various application areas and the immense amount of blogs, the diversity of discussed topics continuously increases. The diversity reaches from travel and news, to politics and gaming.

Blog readers are overwhelmed by the enormous number of blogs and the blogs' diversity. To handle this information overload, the research and application area of blog retrieval evolved [2]. Equally to traditional information retrieval (IR) and data mining approaches, the target is to ease the understanding of the causal relations in the blogosphere and the retrieval of the most relevant blogs to the user's information need [26]. To identify relevant blogs traditional ranking approaches focus on the reach and influence of a blog, but do not consider its content quality.

II. Topic Consistency Rank

We present a ranking approach based on the topical consistency of blogs. This ranking aims to ease the retrieval of expert blogs that are particular important for users to identify blogs to follow and interact with.

We define topical consistency as the degree to which a blog author focuses on a specific set of topics. If blog authors cover several topics, like in random interest blogs or diaries, they have a low topic consistency and thus cannot create topical thrust. In contrast, a blog has the highest topic consistency if it continuously concentrates on one topic. It is argued that such a blog develops a sufficiently high expertise in this topic [3]. Thus, we expect the content of this blog author to be more relevant to an information need than the content of a topically versatile, and influential author. Analogous to frequently cited experts in the real world, we expect that blog readers are more likely to trust and interact with a blog author with a high topic consistency.

We identified four different aspects of topic consistency. The *intra-post* consistency measures the internal consistency of a single post. The *inter-post* consistency defines the consistency among multiple posts of a single blog. The *intra-blog consistency* evaluates the topical consistency between a blog's classification and its posts. Finally, the *inter-blog consistency* measures the consistency of a blog with its linking and linked blogs. The combination of all four aspects results in a rank metric for topic consistency.

In order to evaluate the plausibility of the topic consistency ranking, we formally define and prototypically implement it in the course of this paper. Further, we test whether a correlation between the topical consistency of a blog and its influence is observable. This evaluation can make recourse to a data set that currently consists of 12,000 blogs with over 600,000 posts.

III. Related Work

We divide the related work into three categories of ranking approaches.

The first category consists of *general* rankings that assess web pages and other documents. The second category includes *blog-specific* rankings that are specialized on blogs and other social media channels. The last category comprises *consistency-related* rankings that incorporate the topic consistency of a document or blog into the ranking.

978-1-4799-2902-3/13 $31.00 © 2013 IEEE
DOI 10.1109/WI-IAT.2013.37

A. General Rankings

General rankings like *PageRank* [16] and *HITS* [12] are all based on the web link graph. However, traditional web pages show a different linking behavior than blogs. Blogs offer different types of links, e.g. trackbacks or blogroll links, with different semantics. Furthermore, the blog link graph tends to be rather sparse in comparison to the overall web [10].

B. Blog-Specific Rankings

To address the special characteristics of blogs, blog ranking engines, like *Technorati* and *Spinn3r*, and current research introduce tailor-made ranking algorithms for the blogosphere [5].

Technorati established the *authority* score as their unique ranking. It is calculated based on a blog's linking behavior, categorization and other associated data over a small period of time [23]. Although *Spinn3r* is well known for its crawling service, it also provides a simple *PageRank* and a *Social Media Rank*. The *Social Media Rank* is an adaptation of the *TrustRank* algorithm. It incorporates social networks as incoming link providers and uses a fixed number of initially trusted users to prevent spam. [21]

Current research by Kritikopoulos et al. [13] introduces a ranking score, called *BlogRank*. It is a modified version of the PageRank algorithm. The BlogRank score is based on the link graph and different similarity characteristics of weblogs. The authors create an enriched graph of inter-connected weblogs with additional edges and weights representing the specific features of blogs.

Bross et al. [5] propose the BlogIntelligence-*Impact-Score* (BI-Impact) ranking, a more complete approach to successfully rank blogs. Similar to the above mentioned rankings, they give special weightings for special link types of the blogosphere. They weight the different interaction types of blog authors like links to comments, posts, and to the start page of a blog. In addition, they incorporate among other variables the publishing frequency and the occurrence of trending terms.

Although blog-specific rankings make extensive use of blogs' properties like comments or blogrolls, these ranking do not assess the content quality of a blog or the knowledge of the blog author.

C. Consistency-Related Rankings

Consistency-related rankings are blog rankings that incorporate the topical consistency of a blog. This topical consistency adds to other factors to form one rank for each blog.

A trend detection system, called *Social Media Miner*, is presented by Schirru et al. [19]. They cluster topics for a given period, find relevant terms (or labels), and visualize the term mentions over time as a trend graph. Nevertheless, posts that consistently handle a specific topic have a constant term frequency of topic terms. Thus, topically consistent blogs get a good trend graph, at least for trending topics.

Sriphaew et al. [22] discuss how to find blogs that have great content and are worth to be explored. They show how to identify these blogs, called *cool blogs*, based on three assumptions: *cool blogs* tend to have definite topics, enough posts, and a certain level of consistency among their posts. The authors measure the consistency based on the similarity of topic probabilities of preceding posts.

Among other indicators Weerkamp et al. [25] introduce a topical consistency to improve the effectiveness of topical blog retrieval. Their topical consistency represents the blog's topical fluctuation. The authors define the consistency as a tf*idf-like score over all terms of a blog. Although this measure favors blogs that frequently use rare terms, it does not reflect when a blog author changes the topic from one post to another.

Liuwei et al. [14] describe a spam blog filtering technique that also incorporates the writing consistency of a blog author. Similar to Weerkamp et al., the consistency on topic level is defined as the average topical similarity of posts. The topical similarity is defined as the distance of the posts' tf*idf word vectors. Thereby, blogs with a extremely high topical consistency are expected to be auto-generated.

Another approach for ranking blogs is introduced by Jiyin He et al. [11]. They define a coherence score to measure the topical consistency of a blog. The authors define a consistent blog as a blog that contains lots of coherent posts. A post is coherent to another post if both posts are in the same cluster of the whole collection.

Chen et al. [6] present a blog-specific filtering system that measures topic concentration and variation. They assess the quality of blogs via two main aspects: content depth and breadth. A blog is consistent if it only handles closely related topics. The underlying topic model is derived from Wikipedia[1].

In contrast to related work, the topic consistency rank presented in this paper calculates the consistency of a blog based on multiple aspects. Thereby, it measures the topical consistency at four different granularities and thus offers a differentiated view on the blogs consistency.

IV. Definition and Implementation

To evaluate the topical consistency of a blog author, we define four different facets of consistency that are combined into one ranking.

a) Inter-post consistency: gives the consistency between posts. It investigates whether the contents of the latest posts discuss closely related topics.

b) Intra-post consistency: is the internal consistency of a post. It is a measure that considers to which extend all paragraphs of a post discuss a similar topic.

c) Intra-blog consistency: compares the topic space created by each posts with the topic space created by tags and categories of this post. Therefore, it is a measure for the quality of the blog's classification system.

d) Inter-blog consistency: measures whether a blog is part of a domain expert community. Hereby, the rank of a blog is increased if blogs handling a similar topic link to it. In addition, a blog is boosted if it links to topically related blogs.

[1] http://www.wikipedia.org/

e) Topic consistency rank: combines all four facets into one blog score.

The remainder of this section gives theoretical foundations for each of the underlying partial scores and hightlights essential implementation details. In general, each score implementation consists of a combination of basic SQL constructs like views, permanent and temporary tables.

A. Consistency between Posts (Inter-Post)

As a first step, we formally define the inter-post consistency. The inter-post consistency compares topical distance of succeeding posts. We represent each post as a topic vector. Each component of this topic vector gives the probability of a post talking about one topic. The sum of all vector components is one as usual for a probability distribution.

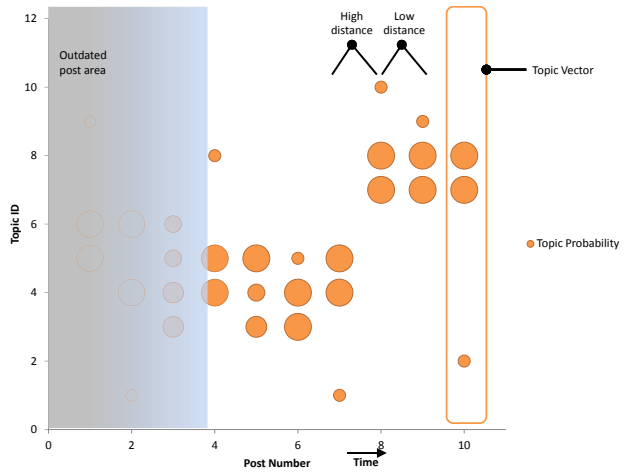

Fig. 1. Visualization of post-topic-probabilities.

Fig. 1 shows the assignment of ten example posts to ten topics. Each column symbolizes a topic vector of a post. The size of a bubble indicates the probability of a post p to be in topic t.

The transient nature of the blogosphere motivates us to only consider the latest posts that lay outside the outdated post area. There are two approaches to define outdated posts: exclude all posts exceeding a specific time span, or including only a specific number of latest posts. The latter solution punishes blogs that are frequently publishing new content by shrinking the observed time window to a day's work. The time span variant is beneficial for small blogs because only a small part of the content is considered. However, we apply the time span variant because we assume that it fits the user's perception.

Sriphaew et al. [22] calculate the average difference of topic vectors of posts with the blog's topic centroid. This favors blogs with a central interest, but does not consider the change of a blog's topic over time. As shown in Fig. 1, blogs can have low distances and high distances between posts. Thus, the average difference of topic vectors of two successive posts serves as indicator for topic consistency.

In the following, we show the formal definition of the inter-post consistency. Before defining the metric, we have to define the sets and functions used for the calculation. The set $Blog$ contains all blogs of the used data set. $Post$ is a set that contains all posts. The set $Post_b$ with $b \in Blog$ contains all posts of blog b. The function $publishedDate(p)$ with $p \in Post$ returns the publishing time and date of a post. The function $LatestPosts_{b,d}$ with $b \in Blog$ and $d \in Date$ being a point in time is a set defined in Eq. 1.

$$LatestPosts_{b,d} = \{p \in Post_b \mid publishedDate(p) \geq d\} \tag{1}$$

$Term$ is the set of all terms. The set $Topic$ contains all topics discussed in the considered subset of the blogosphere. Similarly to Eguchi et al. [8], we define the set $TT_{tp} \subset Term$ as all terms of a topic $tp \in Topic$. All TT_{tp} are pairwise disjoint.

$$\forall tp \in Topic\ \forall j \in Topic : tp \neq j \Rightarrow TT_{tp} \cap TT_j = \emptyset \tag{2}$$

$PT_p \subset Term$ is the set of all used terms of a post $p \in Post$. The function $Prob(p, tp)$ with $p \in Post$ and $tp \in Topic$ gives the probability of the post p being about the the topic tp.

$$Prob(p, tp) = \frac{\sum_{t \in TT_{tp} \cap PT_p} tf * idf(t, p)}{\sum_{t \in PT_p} tf * idf(t, p)} \tag{3}$$

Salton et al. [18] give an overview to the components of the tf*idf-function and its variances. Essentially, it is the product of a term frequency component tf and a collection frequency component idf.

$$tf * idf(t, p) = tf(t, p) \times idf(t, Post) \tag{4}$$

tf is the raw term frequency (number of times a terms occurs in a post). idf is the inverse document frequency. $Post_t$ with $t \in Term$ is the set of all posts in which a term is contained.

$$idf(t, Post) = log \frac{|Post|}{|Post_t|} \tag{5}$$

We define the funtion $topicalDistance(p_i, p_j)$ with $p_i, p_j \in Post$ as the Euclidean distance between the topic vectors of both posts (see Eq. 6). The Euclidean distance is a frequently used distance metric and has proven to apply best for text vector comparison [22].

$$topicalDistance(p_i, p_j) = \sqrt{\sum_{tp \in Topics} (Prob(p_i, tp) - Prob(p_j, tp))^2} \tag{6}$$

The function $predecessor(p) \in Post$ returns the direct predecessor of $p \in Post$. Given these definitions we formalize the inter-post distance as shown in Eq. 7 with $b \in Blog$ and $d \in Date$.

$$interPostDistance(b, d) = \frac{\sum_{p \in LatestPosts_{b,d}} topicalDistance(p, predecessor(p))}{|LatestPosts_{b,d}|} \tag{7}$$

$interPostDistance(b, d)$ is the average topical distance of two succeeding posts among the latest posts of a blog. It returns high values for very inconsistent blogs and low values for very consistent blogs. To give consistent blogs a high inter-post consistency score, we define it as the inverse $interPostDistance(b, d)$, as shown in Eq. 8.

$$interPostConsistency(b, d) = \frac{1}{interPostDistance(b, d)} \tag{8}$$

f) Implementation Notes*:* The inter-post consistency builds upon the tf*idf view based on posts, called *post-tf*idf*, which is also used by the topic clustering (see Sec. V-A). We compute the topic vector differences of each post and its successor. This operation is pretty similar to the intra-post consistency except that it only considers the latest posts.

B. Internal Consistency of Posts (Intra-Post)

The intra-post consistency focuses on the inner consistency of one post. It is high if a blog author focuses on one single topic and does not change the subject while writing one single post. Thus, it favors self-contained and complete posts that do not cover several topics. A consistent post should handle just a few topics, but discuss them in more detail.

The intra-post consistency is very similar to the inter-post consistency except that it operates on the sections of posts. We divide each post into sections by splitting the post's content by each occurrence of more than one line break or HTML separator.

We assign one topic vector to each section. The components of this topic vector represent the probability to which a section is about a specific topic.

We need to define two additional concepts before formalizing the intra-post consistency. Firstly, $Section$ is the set of all sections in the data set and $Section_p \subset Section$ is the set of all sections of one specific post $p \in Post$. Secondly, $predecessor(s)$ with $s \in Section$ is the function that returns the preceding section of one section s.

Further, we define the function $topicalDistance(s_i, s_j)$ with $s_i, s_j \in Section$ in the same manner as Eq. 6.

$$intraPostDistance(p) = \frac{\sum_{s \in Section_p} topicalDistance(s, predecessor(s))}{|Section_p|} \quad (9)$$

We also define the intra-post distance for a whole blog. $intraPostDistance(b, d)$ is the mean of all distance values of the latest posts.

Thereby, we define the $intraPostConsistency(b, d)$ as the inverse intra-post distance to provide consistent blogs with a high score.

g) Implementation Notes*:* To calculate the intra-post consistency, we implement an additional tf*idf calculation view based on paragraphs. We consider all words within a specific window as paragraphs. The size of this window is set to 100 based on the average length of a paragraph, which is 100-150 words [24]. We calculate the topical distance of two succeeding paragraphs and average the results to get first post-level and than the blog-level scores.

C. Consistency between Posts and Classification (Intra-Blog)

The intra-blog consistency serves as a measure for the quality of a blog's classification. It evaluates to which extent the content of posts is consistent with tags and categories that form the *classification system* of a blog. Tags and categories are very important for the orientation of a user and the navigation through the blog. It is crucial that blog authors choose tags and categories wisely and appropriate to their content. In addition, spam blogs tend to overuse tags and categories to earn a higher rank in blog search engines for a high number of keywords. These low quality blogs and spam blogs get a very low intra-blog consistency score.

For a high consistency, tags and categories should span an equal topic distribution as the overall content of a blog. The intra-blog consistency is the distance of the topic vector of each post and the topic vector for the post's classification system.

Before defining the intra-blog consistency we need to formally define define the $Classification_p$ set as the union of categoriy and tag terms of a post p.

Given the classification of each post, $Classification_p$, and the set of all posts in a blog, $Post_b$, we define the intra-blog distance as the average topical distance between each post and its classification (see Eq. 10).

$$intraBlogDistance(b) = \frac{\sum_{p \in Post_b} topicalDistance(Classification_p, p)}{|Post_b|} \quad (10)$$

Finally, we define the $intraBlogConsistency(b)$ as shown in Eq. 11.

$$intraBlogConsistency(b) = \frac{1}{intraBlogDistance(b)} \quad (11)$$

A low value of $intraBlogConsistency(b)$ indicates a mismatch between the classification and the actual content. Thus, the quality of the blog is questionable and it is supposed to be of a lower rank.

h) Implementation Notes*:* We use the *post-tf*idf* view to get the term importance values for the content. Further, we use a tf*idf view based on the classification system, called *class-tf*idf*. This view returns the importance values for each term used in tags and categories. We compute the intra-blog consistency on post-level by the topical distance of the post's classification and the post's content vector and average the distance to get a blog-level score. Blogs get zero intra-blog consistency if they do not use tags or categories.

D. Consistency of Linking and Linked Blogs (Inter-Blog)

Finally, the inter-blog consistency serves as a context-based consistency metric. It measures the consistency between the blog's content and the content of linking and linked blogs. Thus, it measures whether a blog is part of an expert community. An expert community is a set of blogs that discusses on one topic and discuss this topic interactively. For example, during the Arab spring one single blog started the discussion and other blogs built an active discussion around this initial blog [17].

Among other motivations, we identified two targets of the followers of blogs: First, they like to spread the word of the referenced blog author to widen the reach of the message. Second, referencing blog authors want to discuss the message and get into an active discourse with the referenced blog author. Those discourses are the essence of the blogosphere. Similar to Wikipedia, blog authors increase the information quality by evaluating and iterating posts of each other.

Blogs have a set of special link types, but only a few of them are actual interaction links and not only friendly links or advertisements. Blogroll links and links, which are not located in posts or comments, have no evaluating or commenting nature. In contrast, if a blog author links from a post directly to a post of another blog author, the author indicates a reply or similar reaction like a reference. Further, comment authors can also link to other posts, this is formally regarded as a linkback. Linkbacks are also indicators for an active discourse between two blogs. These links, linkbacks and links from posts, are interaction links.

The inter-blog consistency defines the consistency of a blog and blogs that link or are linked via an interaction link.

The post linking post relation (PLP) contains the tuple (p_i, p_j) with $p_i, p_j \in Post$ if p_i has an interaction link to p_j. We define the set IP_{p_i}, incoming posts, with $p_i \in Post$ as follows:

$$IP_{p_i} = \{p_j \mid p_j \in Post \wedge (p_j, p_i) \in PLP\} \tag{12}$$

In parallel, we define the set OP_p, outgoing posts, $p \in Post$.

$$OP_{p_i} = \{j \mid p_j \in Post \wedge (p_i, p_j) \in PLP\} \tag{13}$$

Incoming links cannot be controlled by the blog author. Hence, two constants α, β introduce a weighting for incoming and outgoing posts.

We define the $postContextDistance(p)$ with $p \in Post$ as the weighted sum of the average distance to all incoming and the average distance to all outgoing posts (see Eq. 14).

$$\begin{aligned} &postContextDistance(p) = \\ &\alpha * \frac{\sum_{j \in IP_p} topicalDistance(p,j)}{|IP_p|} + \\ &\beta * \frac{\sum_{j \in OP_p} topicalDistance(p,j)}{|OP_p|} \end{aligned} \tag{14}$$

A typical weighting is $\alpha = 0.6$; $\beta = 0.4$ to slightly emphasize incoming links for their unbiased nature.

We define the $interBlogDistance(b, d)$ with $b \in Blog$ and $d \in Date$ in Eq. 15. The inter-blog distance calculation considers only the latest posts due to the transient nature of the blogosphere.

$$\begin{aligned} &interBlogDistance(b,d) = \\ &\frac{\sum_{p \in LatestPosts_{b,d}} postContextConsistency(p)}{|LatestPosts_{b,d}|} \end{aligned} \tag{15}$$

We define the $interBlogConsistency(b, d)$, analogously to the other three aspects, as the inverse $interBlogDistance(b, d)$ (see Eq. 16).

$$interBlogConsistency(b,d) = \frac{1}{interBlogDistance(b,d)} \tag{16}$$

Implementation Notes We calculate the topical distance between all outgoing and incoming links of a post by joining the post-topic-probability table with the link table (biggest table in data set). By averaging the distances, we compute the blog-level score.

E. Combined Topic Consistency Rank

Finally, we define the topic consistency rank as the combination of all four facets. We combine all facets by calculating a weighted sum for each blog.

We define the $topicConsistency(b, d)$ with $b \in Blog$ and $d \in Date$ in Eq. 17. The four constants, χ, δ, ϵ,and γ, give a weighting for each component of the topic consistency rank.

$$\begin{aligned} topicConsistency(b,d) = &\chi * interPostConsistency(b,d) + \\ &\delta * intraPostConsistency(b,d) + \\ &\epsilon * intraBlogConsistency(b) + \\ &\gamma * interBlogConsistency(b,d) \end{aligned} \tag{17}$$

The weighting can be varied according to the characteristic of the analyzed data set. Caused by the low usage of categories and tags in the *BlogIntelligence* data set and the high usage of content summaries in posts' content, we use the following empirical detected weights: $\chi = 0.3$; $\delta = 0.2$; $\epsilon = 0.2$; $\gamma = 0.3$. For future work we consider the automatic calculation of weights, e.g. by using machine learning techniques.

We calculate the final topic consistency rank by normalizing the results of the $topicCosistency$ function over all considered blogs. Through this normalization the values will be between 0 and 1, which is a common approach for rank normalizations [9].

V. Topic Detection

As mentioned in Sec. IV-A, all topic consistency metrics depend on topic term sets. To find topics and assign terms to topic term sets, we implement a topic detection procedure, shown in Fig. 2.

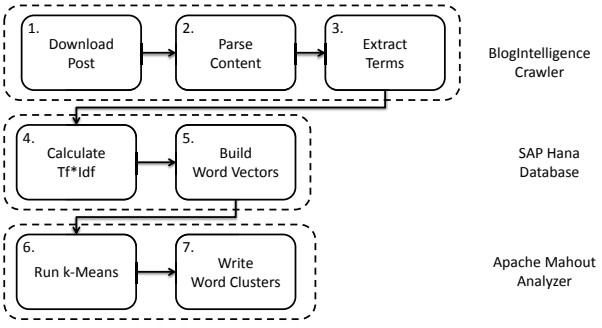

Fig. 2. Flow diagram of the topic detection.

A. Prerequisites

There are several steps necessary before running the actual clustering algorithm, which creates the topic term sets. The preprocessing covers steps 1-5 of the topic detection flow (see Fig. 2).

First of all, we harvest the blogosphere *(Step 1)*, parse the HTML files and remove non-textual content like images and markups. Than we tokenize the texts, remove stop words (stop

word lists from the *Weka*[2] project) and stem the remaining words.

The preprocessing results in a word stem vector for each post that assigns the word stem occurences.

An SQL procedure calculates the tf*idf values for each word *(Step 4)*. The implementation follows Eq. 4. Next, we compute the word vectors for each post that are used for the consistency calculation. Further, we create the post vectors for each word (Step 5) that are necessary for the word clustering.

With step 5 the preprocessing is completed and all vectors can be loaded into the HDFS file system of Mahout[3].

B. Clustering

The two last steps are executed by the Mahout framework. Mahout offers various clustering algorithms like mean shift clustering, spectral clustering, latent Dirichlet allocation, and k-means clustering [15].

As discussed in Sec. IV-A, the topic consistency rank relies on a 1:n relation between words and topics. Thus, we can use the k-means clustering for a simplified prototypical implementation. *k-means* is a well known algorithm for clustering objects that creates pair-wise distinct clusters *(Step 6)*. The choice of a feature vector is crucial for the meaning of the clustering results. By selecting the tf*idf values in each post for each word, we group words together that frequently appear in the same post. Thus, words with a similar meaning are assigned to the same cluster [4].

These word groups are the topic term sets used for the calculation of topical distances. The granularity of the topics is dependent on the user-defined number of clusters *k*. As proposed by Abe et al. [1], the aim is to find clusters with around 100 words per cluster.

VI. Evaluation

This Section discusses the results and the plausibility of the topic consistency rank. Therefore, we show the results of the partial ranks, the overall rank, and compare it to the results of the BI-Impact score.

A. Experimental Setup

For the evaluation of this paper, we activated the *BlogIntelligence* crawler for one month. The crawler uses an 8 core machine with 24 gigabyte RAM running Ubuntu Linux. We store the harvested data in a separate database machine with 32 cores and 1 terabyte RAM running Suse Linux. This machine also runs the SQL analytical queries.

The cluster setup for the topic detection consists of 12 machines with 2 cores and 4 gigabyte RAM each. We group these machines into one Hadoop cluster that is configured to run 50 parallel tasks.

We show the key data indicators of the data set in Tab. I.

The quality of the underlying clustering is crucial for the quality of the topic consistency rank. Especially, the size of clusters determines whether blogs with a versatile interest wrongly get a good consistency rank. As a consequence of our clustering evaluation, the topic consistency rank calculation uses a k-means configuration that creates 18000 clusters with on average 10 words per cluster.

Indicator	Value (approx.)
data set size	500 GB
crawled web pages	2.5 million
identified blogs	12,000
identified posts	600,000
average words per post	57.5
average number of categories per post	2.6
average number of tags per post	4.2
number of news portals	1,300

TABLE I. State of the *BlogIntelligence* data set.

B. Results of the Topic Consistency Sub Ranks

We calculate the ten best blogs for each of the topic consistency sub ranks. We configured the BI crawler to crawl only the German blogosphere. Therefore, the majority of all blogs is German and the top consistency blogs are German, too. For each of the sub ranks, we introduce two highly ranked representatives in detail.

Rank	Intra-Post	Inter-Post
1	promicabana.de	*blog.de.playstation.com*
2	*dsds2011.info*	upload-magazin.de
3	blog.beetlebum.de	blog.studivz.net
4	schockwellenreiter.	der-postillon.com
5	hornoxe.com	*allfacebook.de*
6	netbooknews.de	achgut.com
7	*iphoneblog.de*	gutjahr.biz
8	carta.info	elmastudio.de
9	blog.studivz.net	netzwertig.com
10	seo.at	lawblog.de

TABLE II. The top ten ranked blogs for intra-post and inter-post consistency.

The top ten blogs for the two post-related sub ranks are shown in Tab. II.

One example for a high intra-post consistency is the *dsds2011.info* blog. The intra-post consistency gives the average internal consistency of posts in a blog. *dsds2011.info* is a follower blog of a German TV show that has the aim to cast a new superstar. This blog is a *fan* blog. Therefore, each post mostly focuses on one person, e.g. the current candidate. Further, some posts discuss the performance of each candidate of a show. This causes that each paragraph of such a post focuses on another person, but also uses the same attributes to describe the performance.

Another blog with a high intra-post consistency is the *iphoneblog.de*. Obviously, the topics of each post are all related news about Apple's iPhone. Each post of this blog contains on average five paragraphs, is carefully investigated, and concentrates on one feature, game, or accessory of the iPhone. These special interests are fully investigated in a post over several paragraphs. As a consequence, the internal consistency of the posts is high.

A representative for an high inter-post consistency is the *blog.de.playstation.com* blog. This blog has an high topical

[2]http://www.cs.waikato.ac.nz/ml/weka/

[3]http://mahout.apache.org/

consistency between the latest published posts. The main focus of this blog is on PlayStation games. Hereby, it frequently publishes posts about the latest games, which are discussed regarding their game play, graphics, and story line. Each post presents a game in a similar structure and phrasing. Thus, the topical distance between these posts is very low and the topical consistency is very high.

Another highly ranked blog regarding the consistency between posts is *allfacebook.de*. It publishes posts about new features of the social network, discussion about privacy, and the latest news about Facebook. Although this blogs handles these three topics, it usually publishes multiple posts per topic in a row. This decreases the distance between succeeding posts and boost its inter-post consistency.

Rank	Intra-Blog	Inter-Blog
1	readers-edition.de	*innenaussen.com*
2	*iphoneblog.de*	shopblog author.de
3	eisy.eu	nachdenkseiten.de
4	karrierebibel.de	helmschrott.de
5	meinungs-blog.de	blog.studivz.net
6	dsds2011.info	fanartisch.de
7	*macerkopf.de*	achgut.com
8	kwerfeldein.de	internet-law.de
9	events.ccc.de	*scienceblogs.de*
10	mobiflip.de	events.ccc.de

TABLE III. THE TOP TEN RANKED BLOGS FOR INTRA-BLOG AND INTER-BLOG CONSISTENCY.

The top ten blogs for the two blog-related sub ranks are shown in Tab. III.

One example of an high intra-blog consistency rank is also the *iphoneblog.de* blog. This blog uses the post classification in an appropriate way. As mentioned above, the posts of this blog are carefully edited. By investigating the content of the blog, it is observable that each post contains beside the common categories also at least six content-specific tags. This shows that a blog gains a high consistency ranking for the intra-post and intra-blog consistency by carefully authoring its posts.

Another example is the *macerkopf.de* blog. In contrast to *iphoneblog.de*, the posts of this blog handle a higher variety of topics and comment more critical. For example, they frequently compare the iPhone against other mobile phones. Hereby, a post covers at least two topics. Nevertheless, categories and tags address each topic of the post, which results in a high quality of the classification and in a high intra-blog consistency rank.

The inter-blog consistency measures the consistency of a blog with a linking and linked blogs. The best ranked blog for the inter-blog consistency is the *innenaussen.com* blog. This blog writes reviews about diverse beauty products. The blog link graph indicates that this blog is mainly linking other product reviews e.g. for referencing another opinion on the product. Further, we observe that it is also linked by product review blogs on beauty products like the *lipglossladys.com* blog.

The *scienceblogs.de* blog has also an high inter-blog consistency rank. This is caused by its link directory nature. It mainly collects and summarizes posts from other science-related blogs and provides an entry point into a science community. This blog mainly references the original content. Thereby, its summaries are very consistent with the linked content.

In addition, by comparing all four sub ranks of Tab. II and Tab. III, the *blog.studivz.net* shows high consistency ranks for each subrank except the intra-blog consistency. This blog writes about topics around a German social network called *studiVZ*. It is a typical corporate blog that describes news and new features of a company and the company's products. Hereby, the blog has highly consistent posts that discuss a topic over multiple paragraphs. It constantly posts about activities of the company and is linked by blogs, which spread the news of the company. Nevertheless, each post of this blog is not tagged and is only categorized as *allgemein* (German for miscellaneous), which is a common standard configuration for blog systems.

By investigating the top ten rank blogs for each subrank, we analyzed two examples for each subrank and the evaluation shows that the sub ranks create plausible results.

C. Comparison of BI-Impact and Combined Topic Consistency Rank

The weighted combination of all sub ranks is the combined topic consistency rank. It identifies the topical consistent blogs in the data set. Thereby, it creates a ranking of experts depending on the consistency of their writing. In contrast, the BI-Impact aims to identify the most influential blog authors with the highest reach and famousness.

During the evaluation, we compare both ranks against each other to find possible correlations.

Blog	Combined topic consistency rank	BI-Impact
helmschrott.de	1	85
gedankendeponie.net	2	94
yuccatree.de	3	104
upload-magazin.de	4	96
nachdenkseiten.de	5	117
events.ccc.de	6	54
telemedicus.info	7	118
bei-abriss-aufstand.de	8	90
stereopoly.de	9	87
annalist.noblogs.org	10	88

TABLE IV. TOP TEN RANKED BLOGS FOR THE COMBINED TOPIC CONSISTENCY RANK WITH THEIR BI-IMPACT RANK

First, we investigate the top ten blogs concerning the combined topic consistency rank. As shown in Tab. IV, we list each top ten blog with its ranking position regarding both rankings.

The two sample blogs, *yuccatree.de* and *telemedicus.info*, have high combined topic consistency ranks. *yuccatree.de* has a low inter-post consistency value caused by the diversity of discussed topics. However, it has a high combined consistency score because all remaining three consistency sub ranks are very high. In contrast, the *telemedicus.info* blog focuses only on privacy and patent right discussions. Thus, it has a very high inter-post consistency that results in combination with the proper usage of tags in a high combined topic consistency rank.

The evaluation exemplarily demonstrates the plausibility of the consistency ranking and shows the intuitive inverse relation between topical consistency and popularity in the blogosphere.

VII. Future Work

Future work will cover more advanced topic detection algorithms that might incorporate external resources like Wikipedia, WordNet, or ontologies. One can use the inherent topical hierarchies of these resources to create better clusters that should improve the quality of the topical consistency ranking.

Further, we aim to integrate sentiment detection into the consistency calculation. Thus, the ranking can differ between negative experts or haters that constantly criticize a topic or those who talk positive about it, called fans.

To communicate the results of the topic consistency ranking with users, we will develop a visualization. This will enable users to get a fast overview of the experts and influencers of the blogosphere.

VIII. Conclusion

We proposed a metric for topical consistency of a blog with the goal to identify domain experts in the blogosphere.

To ease the retrieval of these blogs, we defined four different aspects of topic consistency: (1) intra-post, (2) inter-post, (3) intra-blog, and (4) inter-blog consistency. These aspects define the consistency of a blog on different granularities: from the internal consistency of a post's paragraphs to the global consistency between a blog and its linking and linked blogs. We combine the four aspects into a joint rank, called topic consistency rank.

We evaluated the plausibility of the topic consistency rank based on a real world data set. We exemplarily analyzed the top ten results of each aspect and discussed two representatives in detail.

The analysis of the top ten blogs appeared to imply an inverse relation between the topic consistency of a blog and its reach i.e. the more consistent a blog is, the less influence it can gain in the blogosphere. In contrast, by analyzing the distribution of ranks among the top hundred, we cannot observe that there is a correlation between the influence and the consistency of blogs. Thus, we consider both metrics to be independent.

As a consequence, the topic consistency rank is established as an additional indicator, beside the influence of a blog, to ease the retrieval for domain expert blogs.

References

[1] H. Abe and S. Tsumoto. Evaluating a temporal pattern detection method for finding research keys in bibliographical data. pages 1–17, Jan. 2011.

[2] J. Arguello, J. Elsas, J. Callan, and J. Carbonell. Document representation and query expansion models for blog recommendation. In *Proc. of the 2nd Intl. Conf. on Weblogs and Social Media (ICWSM)*, 2008.

[3] K. Balog, L. Azzopardi, and M. De Rijke. Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 43–50. ACM, 2006.

[4] J. Bross. *Understanding and Leveraging the Social Physics of the Blogosphere*. PhD thesis, Hasso-Plattner-Institute, 2011.

[5] J. Bross, K. Richly, M. Kohnen, and C. Meinel. Identifying the top-dogs of the blogosphere. *Social Network Analysis and Mining*, pages 1–15, 2011.

[6] M. Chen and T. Ohta. Using blog content depth and breadth to access and classify blogs. *IJBI*, 2010.

[7] T. Cook and L. Hopkins. Social media or, "how i learned to stop worrying and love communication", September 2007.

[8] K. Eguchi, K. Kuriyama, and N. Kando. Sensitivity of ir systems evaluation to topic difficulty. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, volume 2, pages 585–589. Citeseer, 2002.

[9] M. Fernández, D. Vallet, and P. Castells. Probabilistic score normalization for rank aggregation. *Advances in Information Retrieval*, pages 553–556, 2006.

[10] K. Fujimura, H. Toda, T. Inoue, N. Hiroshima, R. Kataoka, and M. Sugizaki. Blogranger - a multi-faceted blog search engine. WWW, 2006.

[11] J. He, W. Weerkamp, M. Larson, and M. de Rijke. An effective coherence measure to determine topical consistency in user-generated content. *IAPR*, 2009.

[12] J. Kleinberg. Bursty and hierarchical structure in streams. In *ACM SIGKDD*, 2002.

[13] A. Kritikopoulos, M. Sideri, and I. Varlamis. Blogrank: ranking weblogs based on connectivity and similarity features. 2006.

[14] W. Liu, S. Tan, H. Xu, and L. Wang. Splog filtering based on writing consistency. 2008.

[15] S. Owen, R. Anil, T. Dunning, and E. Friedman. Mahout in action. 2011.

[16] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

[17] S. Reese, L. Rutigliano, K. Hyun, and J. Jeong. Mapping the blogosphere professional and citizen-based media in the global news arena. *Journalism*, 8(3):235–261, 2007.

[18] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[19] R. Schirru, D. Obradović, S. Baumann, and P. Wortmann. Domain-specific identification of topics and trends in the blogosphere. *Advances in Data Mining. Applications and Theoretical Aspects*, pages 490–504, 2010.

[20] J. Schmidt. Weblogs: eine kommunikationssoziologische studie. 2006.

[21] Spinn3r. Rank overview, November 2012.

[22] K. Sriphaew, H. Takamura, and M. Okumura. Cool blog identification using topic-based models. IEEE, 2008.

[23] Technorati. What is technorati authority?, September 2012.

[24] J. Tressler, M. Larock, and C. Lewis. *Mastering Effective English*. The Copp Clark., 1980.

[25] W. Weerkamp and M. De Rijke. Credibility improves topical blog post retrieval. ACL, 2008.

[26] I. Witten, E. Frank, and M. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2011.