Seifert, Christin; Witt, Nils; Bayerl, Sebastian; Granitzer, Michael

Article

Digital Library Content in the Social Web: Resource Usage and Content Injection

STCSN-E-Letter

Digital library content in the Social Web: Resource Usage and Content Injection

Christin Seifert
University of Passau
Passau, Germany
Email: christin.seifert@uni-passau.de

Nils Witt
German National Library of Economics
Kiel, Germany
Email: n.witt@zbw.eu

Sebastian Bayerl and Michael Granitzer
University of Passau
Passau, Germany
Email: surname.name@uni-passau.de

***Abstract*—Providers of digital non-mainstream content face two major challenges. First, they have little knowledge about the usage of their content in social media channels, which is neccesary to improve dissemination strategies. The second challenge is how to bring the content to interested users, who do not know their portals. We present two case studies for analysing and injecting content into social media channels. More specifically, we analyse the usage of scientific literature from the domain of economics in blogs and tweets that are related to the field of economics. Additionally, we present two mechanisms for injecting content into blogs or tweets, by the means of a Wordpress plugin and a Twitter bot respectively. The usage analysis shows that the resource coverage is rather low in the investigated social media channels ($\approx 0.15\%$ in tweets and $\approx 0.5\%$ in blogs). Using a Twitter bot for content dissemination is feasible, but not advisable because the automatic approaches are quite hard to optimise towards Twitter's usage policy. For content injection into the blogosphere the Wordpress plugin shows promising results in a qualitative user study. Our results indicate that content injection into social media channels is a promising way for content dissemination and point towards a preference of blogs over twitter as target channel.**

I. Introduction

In the last decade, Europe has made a tremendous effort to bring cultural, educational and scientific resources to the general public. Although such massive amounts of culturally and scientifically rich content are available, the potential of its use for educational and scientific purposes remains largely untapped. One reason can be seen in current web content dissemination mechanisms which are dominated by a small number of large central hubs like major search engines (e.g. Google), social networks (e.g. Facebook) or online encyclopaedias (e.g. Wikipedia). In order to maintain their valuable services, those large hubs have to focus on and are optimised for commercially viable mainstream content. While cultural and scientific resources provide valuable and educational content, they cannot be considered as mainstream. Quite contrary, most of this can be considered as high-quality niche content for a rather small community and forms part of the so-called Long Tail. The Long Tail theory [1], [2], first introduced by Chris Anderson, argues that in internet-based markets, niche content adds up to a huge body of knowledge, but is hidden from most users. In the Long Tail, content is maintained and curated by a large number of small to medium-sized institutions such as memory organisations (e.g. archives and museums), national and digital libraries and open educational repositories. However, the few large Web hubs hardly support the dissemination of this Long Tail content leaving a gap for bringing cultural and scientific wealth into educational and scientific processes.

Additionally, content providers currently lack the means of analysing the usage of their resources in social media channels. If content providers would know in which contexts their content is used, they could devise more effective dissemination strategies and optimise their search or recommendation techniques according to trends in social media channels. For example, a content provider could feature a blog post about some trending topic, which was automatically detected, or automatically rerank resources by boosting resources relevant to current trends in social media.

In this paper we present two case studies for finding resources in social media channels and injecting content into those channels to increase the distribution of resources. Particularly, we focus on scientific papers from the domain of economics and analyse two social media channels, the twittersphere and the blogosphere. More specifically, the contributions of this paper are the following:

- We present an analysis framework for detecting usage of scientific papers with the twittersphere and the blogosphere.
- We present and evaluate a Twitter bot for observing the Twitter stream and proposing scientific resources to economic tweets.
- We present and evaluate a Wordpress plugin for blog writers for adding scientific resources into their blog posts in a cite-as-you-write manner.

The remainder of the paper is structured as follows: Section II presents the Twitter case study, while section III details the case study for blogs. A summary and outlook on future work is given in section IV.

II. Case Study – Twitter

The first case study focuses on the twittersphere, the virtual channel spanned by users and tweets on the popular microblogging platform Twitter [3]. We pursued two goals, first we wanted to investigate the resource usage of economic literature within the Twitter platform, and second we aimed at increasing this resource usage by actively injecting content with a Twitter bot. We report on the usage mining in section II-A and on the development and evaluation of the Twitter bot in section II-B.

A. Twitter Resource Mining

In this section we describe our approach and results for mining usage of economic literature in Twitter.

In Twitter, users can follow other users or are followed by other users. Users post short messages (140 characters), the so-called tweets. Tweets can be shared (retweeted), and contain at least the actual statement, and may additionally contain language information, geo-location, URIs, and hashtags. Hashtags, are a special user-generated markup (e.g., the hashtag "#GenderPayGap"), but need not necessarily refer to the content of the tweet. Twitter provides a publicly available API for accessing its content, but imposes some restrictions, for instance the number of requests per user for a given time interval [4].

1) Implementation: To alleviate the API restrictions, only tweets relevant for the economics domain were collected from the REST API using a list of users who are known to be economists. We compiled the list of users from two sources: (i) a list of economic academics highly active on Twitter, curated by the German Library of Economics[1] (ii) the directory Wefollow[2], listing users with interest in certain topics like e.g. economics or 'finance'. Wefollow is curated by the users themselves and incentively provides them so-called prominence scores. It was used as gold standard by [5]. From the final list we used only 5,000 randomly selected users because of the restriction of the Twitter API. In total, approximately 80,000 tweets per week are collected. Additionally, we used the EconBiz database containing approx. 9 Mio resources from economic literature [6], which forms our set of resources of interest. To identify whether a tweet contains such a resource, we applied simple URI matching by first resolving shortened URIs and then check for an exact match of the URI.

2) Results: The tweets were collected in 2014 over a time period of one month. Table I provides an overview of the statistics. In total 200,000 tweets were collected. Only 0.25% of the tweets contain a link to a resource of interest, i.e. only 500 resources are refered to in tweets, but only 300 (0.15%) of them are unique. Summing up, we found that economic resources can be detected with high precision, but the resource coverage in economic Tweets seems to be rather low ($< 2\%$). Furthermore, the recall of the approach is unknown. Because we apply simple URI matching and do not compare the content of the resources that are referred we might miss some resources.

TABLE I. STATISTICS OF USAGE OF ECONOMIC RESOURCES IN TWITTER

# User seeds	# Tweets	# Resources	# Unique resources
5k	200k	500 (0.25%)	300 (0.15%)

B. Twitter Bot

In this section we describe our approach for injecting economic literature into Twitter using a bot. For automatic injection of resources in Twitter there are three possible strategies: (i) the resource could be included in a reply to an existing tweet, (ii) a Twitter bot might post resources as status updates (which are visible by all its followers), (iii) a bot might respond to a question when his account was mentioned in a tweet. We implemented a Twitter bot (account @RecoRobot and its successor @RecoRobot2) that pursues all three strategies[3]. The Twitter bot is able to recommend resources based on keywords extracted from Twitter status updates using the Twitter Stream and REST API. Further, it responds to queries from users when it is mentioned in a tweet. Also, it actively observes the Twitter stream and detects tweets for which it has relevant resources to propose.

1) Implementation: This section describes the implementation of the Twitter bot to (i) reply to tweets with relevant resources, (ii) respond to user questions with relevant resources.

a) Reply to tweets with relevant resources: A first step to recommend relevant resources to a tweet is to identity whether a given tweet would match with any of the available resources. Second the correct resource has to be found. A tweet was considered to match with any of the resources, if it contained a term or concept contained in a manually curated economic thesaurus. The thesaurus contains about 5,800 concepts with about 32,000 describing terms and is available as Linked-Open-Data [7]. The full set contains terms that are too broad and unspecific, and we therefore filtered the set using the hierarchy information yielding 3,251 terms from the topics "business studies" and "economics". This set was used by the first version of the bot, RecoRobot. The bot listened on the Twitter streaming API and detected tweets containing at least one of the terms. If a term is detected in the content of a tweet content, the tweet's terms are sent as a query to the EconBiz search API, and the most relevant result are returned as a tweet mentioning the original poster. An example is shown in Figure 1.

The second implementation of the bot, RecoRobot2, employs an even more reduced set of keywords. Using a manually constructed blacklist with 54 terms the most frequent false positives are filtered. Additionally, if a tweet is deemed relevant based on its content, the timeline of a user is inspected for additional matching keywords. At least 30 tweets (out of 200 from the users timeline) have to contain keywords from the list to ensure that the user regularly tweets about that topic. Also the tweeting frequency of the new account was reduced to one tweet every 15 minutes.

b) Answering user queries: To actively query the bot for resources, the user has to mention the bot in a tweet. The corresponding tweet content is extracted, and sent to the EconBiz search API after filtering the stop words. An example for a query-based recommendation is shown in Figure 2.

2) Results: While the accounts (RecoRobot and RecoRobot2) were actively tweeting, all processed and produced data, like Twitter user, tweets and recommendations where stored in a local database together with meta information like the current timestamp. Hereby, a dataset with a total size of roughly two gigabytes was accumulated. Key facts like number of retweets, favourites and replies are gathered to evaluate the uptake of resources proposed by the bot.

[1] http://zbw.eu

[2] http://wefollow.com

[3] Source code hosted at purl.org/eexcess/components/twitter-bot

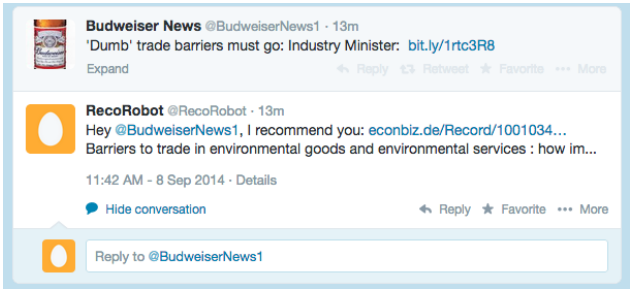

Fig. 1. Twitter bot presenting a resource as response to a tweet detected as relevant.

Fig. 2. Twitter bot offering a resource as response to a tweet detected as relevant.

Table II provides an overview of the statistics for the two deployed versions of the Twitter bot. The first account RecoRobot was running for several months distributing approx. 30,000 recommendations (one recommendation every two minutes). The account was suspended by Twitter and upon requests we did not receive any intelligence of the reason for suspension. We could only guess that the reason might either be the high number of tweets per minute or because the account was reported by some user.

The second account RecoRobot2, which employed a more conservative tweeting behaviour, was active for approx. two weeks and tweeted approx. 800 recommendations. After two weeks, Twitter also suspended this account without further explanation. In its active time approx. 4% of RecoRobot2's recommendations were retweeted. Kwak showed that any retweeted tweet on average reaches 1,000 users independent of the number of followers of the original tweet [3], thus potentially thousands of users could be reached. Also the TwitterBot was able to gather four followers in this period. However, its deactivation shows, that higher quality content together with a lower tweeting frequency is not enough to comply with Twitter's terms of usage.

TABLE II. STATISTICS FOR THE TWO DEPLOYED TWITTER BOTS. COMPLETE STATISTICS ONLY AVAILABLE FOR RECOROBOT2.

RecoRobot	
Total # tweets	approx. 30.000
RecoRobot2	
Total # tweets	798
# Replied	26
# Favourites	18
# Followers	4
# Retweets	6

C. Discussion

The experiments show, that for the economics domain, there is a low coverage of resources in Twitter. While the proposed approach has some limitations, it could still be improved, the reported coverage of resources in economic tweets of approx. 2% is a good estimate. The Twitter bot showed that although injecting resources pro-actively in Twitter is feasible, the effort-gain ratio is bad. This is mostly due to the complex optimization of the automatic approach in order to adhere to the (possibly changing) Twitter policy.

III. CASE STUDY – BLOGOSPHERE

In this section we will introduce the second case study, that is made up of two pieces. First, there is a blog crawler, that is working towards the goal of identifying EconBiz[4] contents in blogs with an economic focus. Secondly, there is a plug-in for Wordpress[5] that aims at injecting economic resources into blogs.

A. Blog Resource Mining

The following part depicts the concept and the implementation of the blog crawler and outlines the results.

1) Approach: The goal was to implement a mechanism that identifies resources (like topics, persons, publications) in blog posts and regains those (and similar) resources in EconBiz. The result is a mapping between blog entries and EconBiz resources. The resources meta data could then, for instance, be used to enhance recommendation services by recommending scientific papers or other blog entries for a given blog entry.

a) Blog Crawler: To tackle that goal, we identify the economists' blogs with the highest impact. As it turned out, there were some sources that already dealt with that issue. We've chosen a list by onalytica.com[6] as a source for influential blogs. We began with developing a generic crawler (i.e. a single crawler, capable of crawling different websites) as proposed in [8], but, since not enough common patters (in terms of DOM-structure) could be found in the hand-selected websites, we had to withdraw this strategy, and, instead, create tailor-made crawlers for each website[7] (focused crawler). The approach discussed in [8] acts on the assumption, that each blog is hosted either by wordpress.com or blogsport.com, which leads to vastly uniform DOM-structures. In contrast, the blogs that we examined are all self-hosted and highly customized.

The main benefit of the focused crawlers is, that they work more precisely and gather more information (e.g. user comments and author) that would have been omitted by the broad crawler. But this approach scales poorly, due to the fact that each new website requires its own crawler.

[4]EconBiz is ZBW's search portal for economics
[5]http://wordpress.org
[6]http://www.onalytica.com/blog/posts/top-200-influential-economics-blogs-aug-2013 retrieved in June 2014
[7]Source code hosted at purl.org/eexcess/components/research/blogcrawler

The blog crawler is based on scrapy[8], a python framework to facilitate the development of web crawling applications. For persistence purposes we use Elasticsearch[9]. Scrapy alleviates the task of web crawling vastly. Essentially, the programmer only has to specify the DOM-nodes that are to be extracted using XPath. The following example illustrates that:

```
sel.xpath(
    '//a[@rel="author"]/text()')
```

When the aforementioned code is applied to a HTML-document containing this HTML-tag:

```
<a rel="author"]
ref="<some URL>">John Doe</a>
```

the code snipped returns `"John Doe"`.

b) Data Analysis: We started with the naive and straightforward approach of URL-matching. Basically, we were looking for URL-patterns that match EconBiz URLs (i.e. `http://www.econbiz.[de|eu]/Record/<title>/<ID>`).

But we also employed a more sophisticated approach [10]. Due to the fact that EconBiz stores bibliographical information (such as author, title, publisher etc.) of scientific papers, we decided to look for URLs linking to PDF-files, as they usually have an author and a title (and some have even more meta information that can be found in Econbiz), assuming the examined file is a scientific paper. Afterwards we have implemented the following strategy for each document in the corpus:

1) The program checks whether the document's meta data field *author* and *title* are not empty. If so, it sends a query assembled from these strings via string concatenation to EconBiz. Then the result list is inspected. In case the result list is not empty, all results are examined and assessed (details are outlined in the next paragraph). When there is no result above a predefined quality threshold, the second stage is executed, otherwise the result is stored and the processing continues with the next document.
2) After the text of the first page of the file is retrieved, the text is divided into smaller chunks (using punctuation and whitespace symbols). The processing of these parts of sentences is identical to the processing of the meta data in the previous step. If there is no result satisfying the quality threshold no match was found, otherwise the list of potential matches is stored.

To assess the quality of the results, the fuzzy string comparison library (fuzzywuzzy[11]) is employed. It contains a method that is invoked with an arbitrary string (selector) and a list of strings (choices). The method returns the choices sorted by how well they match the selector. Each item of the list is accompanied by a value n (where $n \in [0, 1]$), which serves as a quality indication (1 perfect match, 0 no match). The following example illustrates that:

> Given the following names:
> `Barbara Bergmann`
> `C. Fred Bergsten`
> `Frank J. Chaloupka`
> which one matches *Bergmann* best?
> Fuzzywuzzys answer looks like this:
> `Barbara Bergmann 0.90`
> `C. Fred Bergsten 0.55`
> `Frank J. Chaloupka 0.33`
> Unsurprisingly, *Barbara Bergmann* matches best. *C. Fred Bergsten*s score is better because his name contains the substring *Berg*. *Frank J. Chaloupka*s scores worst because there are no notable common patterns.

The quality indicator is used to decide whether a document matches the search query. Qualitative experiments with randomly selected documents have shown, that, if $n < 0.8$ the document and the EconBiz result most likely do not match. But there are still false positive cases with $n \geq 0.8$. Not until $n > 0.9$ the false positive cases are negligible.

For text extraction from PDF-files we used the python library pdfminer [12]. We limited the extraction to the first page, because this task was time-consuming ($\sim$1 second per page). Furthermore, we assume that the first page contains the authors name and the title of the document, which is true for most scientific papers (unless there is a cover page).

2) Results and Discussion: Since the development of a focused crawler for each blog requires a lot of effort, we decided to limit the amount of blogs to 10. Because we wanted to investigate the most influential blogs, we have chosen the top ten websites (with one exception) according to the aforementioned list by Onalytica. Technical limitations (i.e. authentication required) prevented the analysis of one website. Furthermore, we limited the blog crawler to blog entries that had been published in the last year. That yielded approximately 80k entries.

We discovered, that there are no EconBiz URLs in the corpus. This could be due to the selection of the blogs, the majority of which are from the United States. Therefore the URL-matching approach was discarded and replaced by a more sophisticated approach based on meta data recognition in PDF-Files. A subset of 100 files (randomly chosen) of the 2,000 PDF files, that had been found in the corpus, have been investigated manually. Roughly 15% of the documents could be found on EconBiz (and are therefore assumed to be economic papers). The other 85% consist of marketing material, research papers that can be assigned to different (i.e. non-economic) research fields (e.g. human medicine), court decisions, authority documents, etc..

The automatic identification process described in III-B1 can only identify 2%. The gold standard (15%) shows that the current approach is insufficient and needs to be improved.

The current implementation of the data analyzer requires too much computational resources (i.e. CPU-time and mem-

[8] http://scrapy.org
[9] http://www.elasticsearch.org/
[10] Source code hosted at purl.org/eexcess/components/research/bloganalyzer
[11] https://pypi.python.org/pypi/fuzzywuzzy/0.2
[12] https://pypi.python.org/pypi/pdfminer/

ory) for a single machine. The recognition of a single document takes several seconds. Therefore, either optimization or adaption to a cluster of machines is essential in order to make the service beneficial in a real-world application.

Table III highlights the numbers that have been covered in this section.

TABLE III. RESULTS FOR THE BLOGOSPHERE CASE STUDY

Blog crawler	
# Blog entries	80k
# Blogs	10
Gold standard subset	100 papers
Economic papers in gold standard	15%
Automatic identification	2%
Resource coverage	0,5%

B. *Wordpress Plug-in*

This section presents a Plug-in for the Wordpress whose goal is to foster scientifically written blogs. Moreover, we present the results of qualitative user evaluation.

1) Prototype Description: In order to bring more scientific content into blogs we developed a plug-in for Wordpress, that allows the user exploring, citing and embedding EconBiz-Resources. We decided to write a plug-in for Wordpress, because it is widely used[13] and has a big and lively community. The target group for the plug-in are researchers in the field of economics. The plug-in allows blog authors to create scientific citations in a cite-as-you-write manner and thus, make their blog entries appear in a more scientific fashion. Currently the plug-in allows the user to search for Econbiz objects and integrate them into the text without switching the context (i.e. switching website or application). In Figure 4 the editing area (1) and the plug-ins' result list (2) in the back end are shown. Figure 3) demonstrates the result that is presented to the blogs' visitors on the front end.

To give deeper insights into the plug-in, two aspects are investigated in more detail:

1) **Single-page application**
   In order to get a smooth user-experience, the plug-in was designed following the principles of a single-page application (SPA), including the lack of page-reloads after user interaction. To achieve this, data is requested asynchronously via AJAX and, instead of reloading the page each time new data arrives, the websites' DOM is manipulated on-the-fly by means of the jQuery library[14]. Furthermore, the application was built to work in a concurrent manner and is driven by events, which is the foundation for a SPA.
2) **Citation engine**
   The plug-in includes a feature that allows the user to automatically generate citations for EEXCESS resources, that appear in the result list just by clicking a single button. This feature is based on citeproc-JS[15], a library to generate citations with different styles based on templates. The programmer has to provide bibliographic information (such as author, title, release date etc.) and a template for the desired citation style. The library then generates an appropriate citation string.
   The templates are written in a domain specific language called Citation Style Language (CSL)[16]. There are currently more than 7.000[17] styles available that can be easily integrated into the plug-in, which makes this approach very flexible.

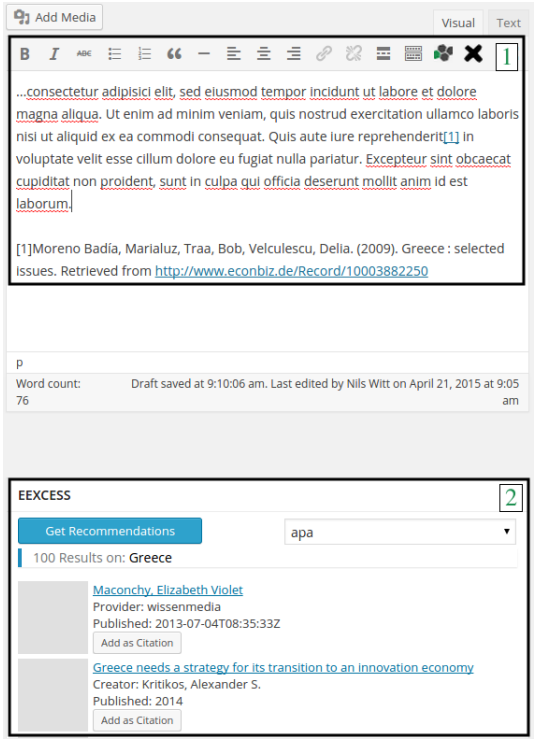

Fig. 3. Plug-in in the Wordpress back end

Fig. 4. Corresponding blog post

2) User Evaluation: We conducted a qualitative user evaluation based on an early prototype of the plug-in, as we wanted to estimate the usefulness and user acceptance of a tool supporting blog authors by recommending content that can be easily integrated into a blog post.

The five participants are all economists and bloggers using Wordpress on a regular basis. After a test environment (stan-

[13] http://w3techs.com/technologies/overview/content_management/all
[14] http://jquery.org
[15] https://bitbucket.org/fbennett/citeproc-js/overview
[16] http://citationstyles.org/
[17] https://www.zotero.org/styles

dard Wordpress installation with installed plug-in) had been set-up, the participants were invited to create new blog entries using the test environment for two weeks. They were asked to create at least five entries using the features provided by the plug-in. Even though the test environment was accessible over the internet, no audience was invited, because the blog author's feedback was the focus of this evaluation. After two weeks had passed, the participants requested to fill a questionnaire. Beside the question whether or not such a plug-in is useful for blogging, the questionnaire was focused on discovering new features the target group wants, in order to make the plug-in more useful.

All participants judged the plug-in as 'useful' or 'principally useful', but they also noted that improvements are necessary. Several ideas have been mentioned, the most important are:

- More citation styles are needed (the test environment offered five).
- The plug-in should not only cite scientific papers but also embed images into blog posts.
- Sorting and filtering mechanisms are needed, in order to explore the result list effectively.
- Clauses informing about the legal situation of the recommended resources (in particular the images) are desired, to avoid legal conflicts.

C. Discussion

The blog crawler shows, that it is reasonable to gather a small and domain-specific data corpus. Moreover, the experiments, conducted on that corpus, indicate a significant usage of economic papers in the economic blogosphere. Nevertheless, the algorithms used for matching meta data to the respective resource lack high precision. The overall positive feedback obtained in the user evaluation reveals general interest in cite-as-you-write tools that ease the process of creating blog posts which encourages us to further advance the development of the Wordpress plug-in. By now, new citation styles have been added and support for embedding images has been introduced. The remaining ideas are currently being implemented.

IV. Summary and Outlook

In this paper we presented two case studies, in which we first analysed resource coverage of Long Tail resources in social media channels, and second injected resources into those channels. We focused on resources from the economics domain and used a controlled vocabulary from this domain in the Twitter use case. The blogs use case does not depend on such a vocabulary, thus we expect the technology to be applicable to other scientific domains as well. For the two investigated social media channels, twittersphere and blogosphere, we found a low resource coverage. Although both injection technologies, the Twitter bot and the Wordpress plug-in, were found to be feasible, implementing a Twitter bot for dissemination seems generally not advisable because it is hard to adapt the parameters to the (changing) terms of use imposed by Twitter. Consequently, in future work we will focus on the Wordpress plug-in which will be improved using the results of the qualitative user study and re-evaluated again.

We expect to increase the significance of the blogosphere case study, enlargement of the current data corpus is required. The methods described by Berger et al. [9] are a reasonable starting point. In addition, following the argumentation of [10] the blogs with the highest impact cover more domains. Therefore, it is worthwhile to consider smaller but more specific sub-group within economics to obtain more accurate results.

Acknowledgment

The presented work was developed within the EEXCESS project funded by the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement number 600601.

References

[1] C. Anderson, "The long tail," *Wired*, vol. 12, no. 10, October 2004. [Online]. Available: http://www.wired.com/wired/archive/12.10/tail_pr.html

[2] A.-L. Barabási, R. Albert, and H. Jeong, "Scale-free characteristics of random networks: the topology of the world-wide web," *Physica A: Statistical Mechanics and its Applications*, vol. 281, no. 1–4, pp. 69 – 77, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378437100000182

[3] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 591–600. [Online]. Available: http://doi.acm.org/10.1145/1772690.1772751

[4] M. Boanjak, E. Oliveira, J. Martins, E. Mendes Rodrigues, and L. Sarmento, "Twitterecho: A distributed focused crawler to support open research with twitter data," in *Proceedings of the 21st International Conference Companion on World Wide Web*, ser. WWW '12 Companion. New York, NY, USA: ACM, 2012, pp. 1233–1240. [Online]. Available: http://doi.acm.org/10.1145/2187980.2188266

[5] M. Pennacchiotti and A.-M. Popescu, "Democrats, republicans and starbucks afficionados: User classification in twitter," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '11. New York, NY, USA: ACM, 2011, pp. 430–438. [Online]. Available: http://doi.acm.org/10.1145/2020408.2020477

[6] T. Pianos, "Econbiz – meeting user needs with new technology," *LIBER Quarterly*, vol. 20, no. 1, 2010. [Online]. Available: http://liber.library.uu.nl/index.php/lq/article/view/URN%3ANBN%3ANL%3AUI%3A10-1-113576

[7] J. Neubert, "Bringing the "Thesaurus for Economics" on to the Web of Linked Data." in *LDOW*, ser. CEUR Workshop Proceedings, C. Bizer, T. Heath, T. Berners-Lee, and K. Idehen, Eds., vol. 538. CEUR-WS.org, 2009. [Online]. Available: http://dblp.uni-trier.de/db/conf/www/ldow2009.html#Neubert09

[8] G. Petasis and D. Petasis, "Blogbuster: A tool for extracting corpora from the blogosphere." in *LREC*, 2010.

[9] P. Berger, P. Hennig, J. Bross, and C. Meinel, "Mapping the blogosphere–towards a universal and scalable blog-crawler," in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, Oct 2011, pp. 672–677.

[10] P. Berger, P. Hennig, and C. Meinel, "Identifying domain experts in the blogosphere – ranking blogs based on topic consistency," in *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on*, vol. 1, Nov 2013, pp. 252–259.

[11] D. Besagni and A. Belaïd, "Citation recognition for scientific publications in digital libraries," in *Document Image Analysis for Libraries, 2004. Proceedings. First International Workshop on*. IEEE, 2004, pp. 244–252.