Predicting the Evolution of Taxonomy Restructuring in Collective Web Catalogues

Natalia Prytkova, Marc Spaniol, and Gerhard Weikum
Max-Planck-Institut für Informatik
Campus E1 4
Saarbrücken, Germany
{natalia|mspaniol|weikum}@mpi-inf.mpg.de

ABSTRACT

Collectively maintained Web catalogues organize links to interesting Web sites into topic hierarchies, based on community input and editorial decisions. These taxonomic systems reflect the interests and diversity of ongoing societal discourses. Catalogues evolve by adding new topics, splitting topics, and other restructuring, in order to capture newly emerging concepts of long-lasting interest. In this paper, we investigate these changes in taxonomies and develop models for predicting such structural changes. Our approach identifies newly emerging latent concepts by analyzing news articles (or social media), by means of a temporal term relatedness graph. We predict the addition of new topics to the catalogue based on statistical measures associated with the identified latent concepts. Experiments with a large news archive corpus demonstrate the high precision of our method, and its suitability for Web-scale application.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Experimentation, Measurement

Keywords

Concept Mining, Taxonomy Evolution, Content Classification

1. INTRODUCTION

The constantly evolving Web reflects the evolution of society in the cyberspace. Projects like the Open Directory Project (`dmoz.org`) or Yahoo Directory (`dir.yahoo.com`) can be understood as a collective memory of society on the Web. It represents the topical knowledge in a structured way, in form of taxonomy. This taxonomy consists of topics, which are connected by "parent-child" relations. A parent topic is broader, whereas all its children deal with more narrow topics. Each topic in the taxonomy holds a set of links to related resources in the Internet. The taxonomy, once constructed, is constantly evolving, reflecting the human cognition of societal trends. There are several types of changes which can be performed in the taxonomy: topic merging or splitting, renaming, removal, and addition. In this paper we focus on the formation of new concepts that lead to adding a topic to the taxonomy.

New concepts first appear, in latent form, in sources external to the Web catalogues: in news, blogs, and other social media. Figure 1 shows the presence of the words "Japan nuclear plant tsunami" in news articles as a function of time. There is a considerable and sudden increase of interest in these terms around March 2011. The intensity remains at a high-level for several months. The bottom part of Figure 1 shows two consecutive versions of the DMOZ topic "Safety and Accidents". The snapshots of this topic from late 2010 and mid 2011 differ in the subtopics they contain. There is a clear correlation between the massive emergence of news dedicated to the news coverage about "Japan nuclear plant tsunami" and the extension of the DMOZ taxonomy with the topic "Fukushima 2011".

Temporal analysis of news archives reveals that concepts discussed in news media behave differently over time. Some concepts appear occasional while others are persistent. Occasional concepts attract the attention of news media for very short periods of time and fade out afterwards. An example is a meeting of Germany's chancellor with the American president. Concepts of this kind do not lead to new topics in the Web catalogue. Persistent concepts, on the other hand, receive attention almost every day over long time periods; an example is global warming.

In this paper we aim to predict such long-living concepts at their very onset when the number of relevant news articles rises rapidly. These concepts, e.g. "Fukushima 2011", are those that should lead to a novel topic in the Web catalogue. Our goal not only is semantically challenging, but also entails scalability problems. Catalogues like DMOZ contain hundred thousands of topics whose evolution we would like to predict based on 100,000's of daily news articles (not even counting social media). Predictions can be an asset for informing Web catalogue editors about emerging concepts.

We have developed a data-analysis and prediction system called PIWO (Predicting evolution In Web catalOgues) PIWO has the following salient properties: 1) models for extracting emerging latent concepts and predicting novel topics in a taxonomy, based on a temporal term relatedness graph built from news articles; 2) a judiciously designed clustering algorithm, based on maximal cliques in the temporal term relatedness graph, to identify latent concepts in a scalable manner using Map-Reduce computing; 3) an algorithm for predicting structural changes in Web catalogues, based on statistical measures of latent concepts; 4) experiments with the New York Times (NYT) archive for concept extraction

Copyright is held by the author/owner(s).
Fifteenth International Workshop on the Web and Databases (WebDB 2012), May 20, 2012 - Scottsdale, AZ, USA.

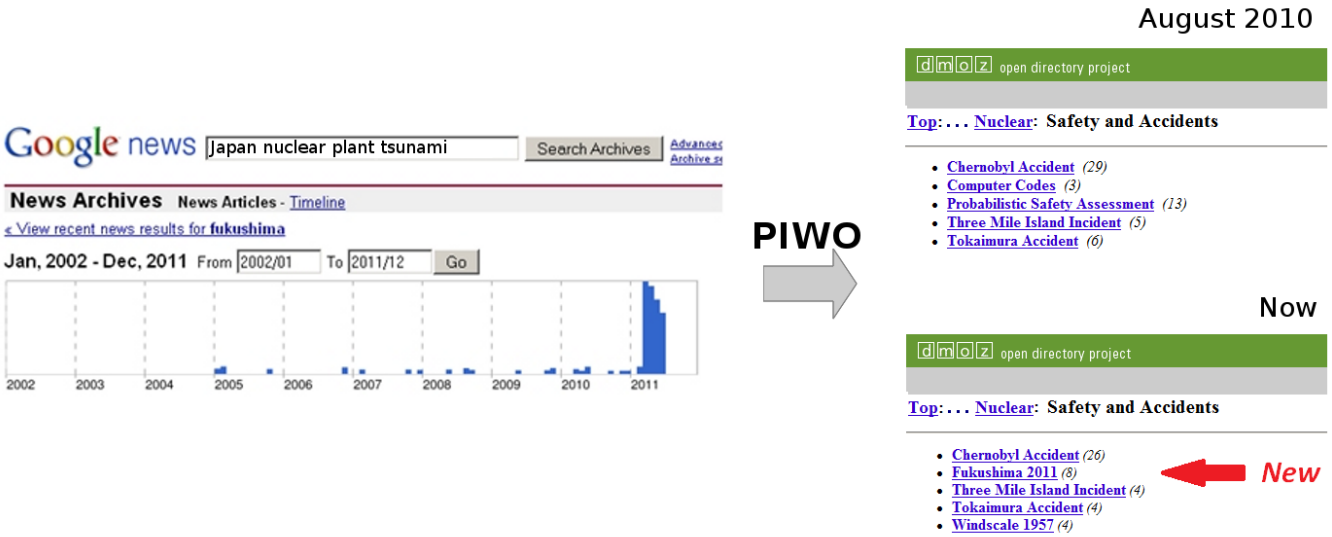

Figure 1: "Japan nuclear plant tsunami" in the news and the emergence of the new DMOZ topic "Fukushima 2011" in "Safety and Accidents"

and predicting new sub-topics in the "Health", "Business" and "Science/Technology" topics of the Open Directory Project (`dmoz.org`). We believe that PIWO is a valuable tool for maintaining Web catalogues, by easing the task of editors and strengthening the quality of crowdsourcing-based methods.

2. CONCEPT MINING

The input to the PIWO system is a time-ordered sequence of *snapshots* of news and taxonomy states. Each news snapshot is either a batch of newspaper articles from an archive (e.g., on a monthly or weekly basis) or a set of incoming articles from online news and social media (e.g., on a daily basis). Each taxonomy state is a tree or DAG of explicitly named topics. Typically, the taxonomy changes on a monthly basis. Figure 2 shows the architecture of PIWO. It consists of two main components: the *concept miner* and the *taxonomy change predictor*. The concept miner discovers latent concepts in the news snapshots, which can enrich the next state of the taxonomy. The taxonomy change predictor takes as input a set of emerging concepts, compares them to the current state of the taxonomy, and identifies topics expected to be restructured by adding a new sub-topic.

2.1 Concept Candidate Gathering

To gather latent concepts we employ clustering on terms (words or phrases), using co-occurrence-based relatedness measures as similarity. This should allow term polysemy, terms not belonging to any concept (noise), and a variable number of latent concepts. These goals determine the class of appropriate clustering methods: hierarchical agglomerative clustering (HAC) allowing overlaps between clusters as well as left-out data points (not assigned to any cluster) [16]. The relatedness between two terms can be computed by the cosine similarity of the corresponding document vectors. The relatedness of terms has a temporal nature and is defined with respect to a timepoint. For instance, shortly after a presidential election, the terms *president*, *elections*, *results* may form a concept and be closely related to each other, while a year after the election they may not co-occur anymore.

Based on the pairwise term relatedness, the sets of mutually related terms at time t can be extracted. One way of modelling the inter-related term sets is to represent them in form of a graph. Nodes of the graph correspond to the terms from the lexicon W. If the relatedness of two terms is above a predefined threshold, the corresponding nodes are connected with an edge.

Definition An undirected graph $TRG^t = (V, E)$ is a *term-relatedness graph* at time t, where V is the set of terms in the lexicon W at time t and an edge $e_{i,j} \in E$ exists if $\forall i, j \in V$, $i \neq j : rel(w_i, w_j, t) > \tau_{rel}$. τ_{rel} is a specified threshold value for term relatedness. $rel(w_i, w_j, t)$ is the cosine similarity between document vectors corresponding to terms w_i, w_j at time t.

The term sets identified this way are referred to as call *concept candidates*. The set of candidates is constructed for each snapshot of the underlying news data. The candidates can then be temporally ordered by their first appearance.

Definition Let W be the lexicon of the snapshot at time t. A set of terms $c^t = \{w_1, w_2 \ldots w_n\} \subseteq W$ is a *concept candidate* at time t, if $\forall\ i, j \in [1..n]\ i \neq j : rel(w_i, w_j, t) > \tau_{rel} \geq 0$. The set of candidates for time t is denoted as C^t.

Our clustering method on the term-relatedness graph can be seen as a horizontal cut in a HAC dendrogram. Given the threshold τ_{rel}, the clustering is then equivalent to finding all maximal cliques in the graph. As this is an expensive computation, we devised a distributed Map-Reduce algorithm to parallelize and speed up the clustering (inspired by Wu et al. [19]). To this end, the graph is partitioned with replication. We hash-partition the graph nodes and store all first- and second-order neighbors of the respective nodes in the same partition. Then each partition computes maximal cliques in parallel. By the definition of a maximal clique,

no clique can be missed this way nor will there be any false positives. Of course, this parallelization does not escape the NP-hardness of the problem: a huge clique will slow down one of the partitions and will be the bottleneck in the Map-Reduce computation. Empirically, however, we found that most cliques are not that big, so that we achieved good scale-up performance. This clustering is carried out once for each snapshot; so our scalable method is crucial for practical viability.

2.2 Concept Candidate Dynamics

For terms sets to be truly relevant concepts they have to present for an extended time period, spanning multiple snapshots. However, such a set of co-occurring terms may vary over time, dropping and adding terms across snapshots. If this variation is low, we refer to the latent concept candidate as a *stable term set*. The deviation of two candidates $dev(c^t, c^{t'})$ is measured by the Jaccard similarity of the corresponding term sets.

Definition Let $\tilde{c} = \{c^{t_j}, \dots c^{t_{j+s}}\}$ be a sequence of concept candidates for the time window from t_j to t_{j+s}. We call $\tilde{c}$ a *stable term set* for time point t_{j+s} if$\forall\ i \in [j \dots j+s-1]:\ dev(c^{t_{j+s}}, c^{t_i}) \leq \tau_{dev}$, where τ_{dev} is a specified threshold for the deviation.

For computing stable term sets, we scan the concept candidates in time order using a window of size s snapshots, and compare term sets between successive snapshots. For the qualifying candidates, we can then express the *concept dynamics* as a sequence of occurrence counters.

Definition The *dynamics of concept* $\tilde{c}$ is defined as $dyn(\tilde{c}) = \{occ(c^{t_i}), \dots occ(c^{t_{i+s}})\}$ where $occ(c^t)$ is the number of articles containing candidate c^t.

We are now ready to define the appearance of a new concept, emerging in a time window but not present at the begin of the window.

Definition The concept dynamics $dyn(\tilde{c})$ forms an *emerging pattern* if there is a time point $t_{i+k} \in [t_i \dots t_{i+s}]$ such that:

$$\forall t' < t_{i+k}\ occ(c^{t'}) = 0 \wedge\ \forall t' \geq t_{i+k}\ occ(c^{t'}) > 0$$

2.3 Concept Candidate Cognition Level

Among all candidates with stable term sets and emerging patterns, we select only those with high *cognition level*, as measured by the frequency and prominence of candidate occurrences. This measure considers the number of documents containing the term set, the lifespan of the concept, and the quality of the sources where we observe the concept within a time window. The source quality is relevant if our news snapshots comprise different collections, for instance, newspapers, blogs, and social-media postings.

Definition Let R be the set of news collections. The *cognition level* of concept c reached within the time window $[t_i \dots t_{i+s}]$ is the aggregation of cognition levels reached in each collection within this sliding window:

$$cogn_{t_{i+s}}(dyn_R(c)) = \sum_{r \in R} quality(r) \cdot \frac{docs_r(c)}{lifespan_r(c)}$$

where $quality(r)$ is a measure for collection quality, $lifespan_r(c)$ denotes the number of snapshots in $[t_i \dots t_{i+s}]$ in which c exists in collection r, and $docs_r(c)$ is the total number of documents in r containing c within $[t_i \dots t_{i+s}]$.

This definition gives flexibility in treating information from different sources. In our experiments, the collection solely consists of the NYT news archive; we will set $quality(r) = 1$ in this special case.

2.4 Interesting Concepts

We now define a notion of *interesting concept* (inspired by [18]), by combining the following three requirements: 1) a concept must be a stable term set for some time window in the news history, as opposed to a sporadic, very short-lived concept; 2) it must be an emerging pattern at some time point (not necessarily in the window of term-set stability), as opposed to a persistent concept that exists independent of time; 3) it must reach sufficient cognition level across all snapshots, as opposed to a highly special, low-caliber concept.

3. TAXONOMY CHANGE PREDICTION

Given the formation of new latent concepts, the task now is to predict a new sub-topic that will be added to the taxonomy of a Web catalogue. The *topics* in a taxonomy are hierarchically organized by means of several relations, most notably, a specialization/generalization tree or DAG. Another relation connects related topics, orthogonally to the main hierarchy. In the following we focus on this main hierarchy, assuming that it is a DAG. Each topic node is associated with a set of *Web links* and short text snippets describing the corresponding Web sites that have been assigned to the topic.

Definition A *taxonomy* at time point t is a directed acyclic graph $T^t = \langle N, E \rangle$, where the node set N corresponds to a set K^t of topics and the edge set E corresponds to topic pairs $k_1, k_2 \in K^T$ such that k_2 is thematically subsumed by (i.e., is more narrow than) k_1. We assume the graph has a single root $Top \in N$ that comprises all topics in the taxonomy. Each node k is associated with a set of terms $Ext(k)$ consisting of the terms in the topic description and the descriptions and URLs of the topic's Web links. $W_T^t = \cup_k Ext(k)$ is the lexicon of T^t.

When new substantial concepts appear in the news or in the blogosphere, a taxonomy tends to capture these changes, by adding a new sub-topic under an existing parent topic.

We hypothesize that all interesting concepts, as identified by the methods in Section 2, should lead to new topics in the next snapshot of the taxonomy. As observed in [4], sometimes not a single topic, but an entire subtree is inserted. This happens when a new topic contains a set of auxiliary topics. For instance, the topic *'Authors'* is likely to be added along with an alphabetical list to search an author by last name. We do not aim at predicting auxiliary changes of that type.

Definition Let T^t and T^{t+1} be consecutive snapshots of a taxonomy at time point t and $t+1$ respectively. The *creation of a new concept* is reported, iff there is a topic k' present in T^{t+1} but not in T^t and a topic k present in both T^{t+1} and T^t such that (k, k') is an edge in T^{t+1} and the extension of k does not overlap by more than θ with any topic in T^t, where θ is a specified threshold.

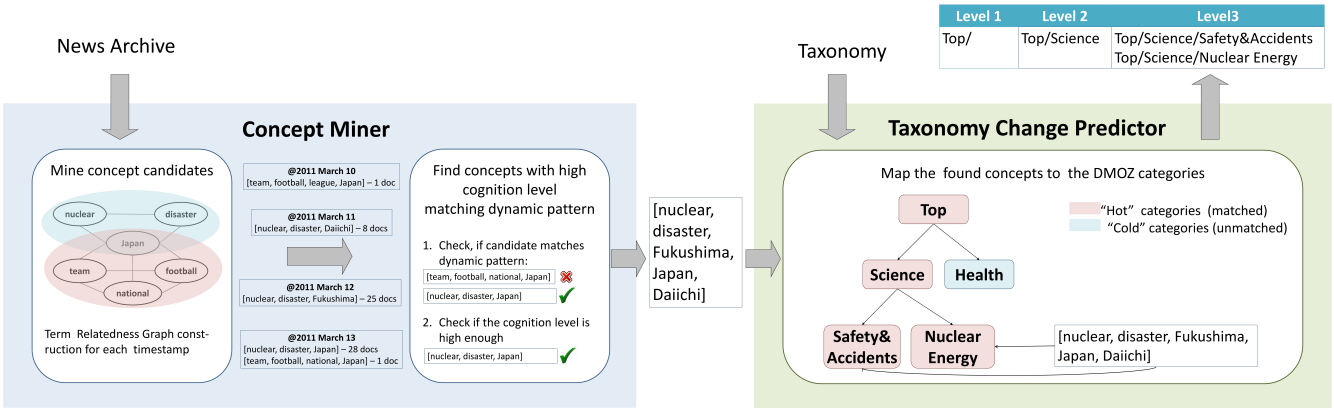

Figure 2: Overview of the PIWO system

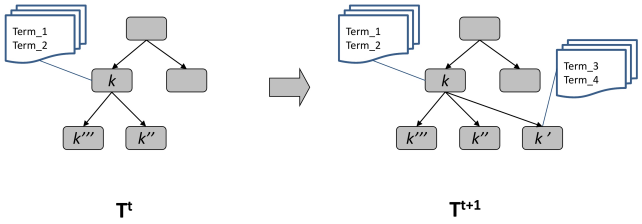

Figure 3: Creation of a new concept

The property that a new topic k' does not inherit a large portion of Web links from any pre-existing topic reflects the freshness of k'. Otherwise, k' could have been derived from an old topic by renaming or merging/splitting categories. Such purely structural changes are not considered here.

In order to predict taxonomy changes, we devise two strategies: concept-based and topic-based prediction.

3.1 Concept-based Prediction

This type of prediction is based on the concept dynamics and the concept cognition level discussed in the previous section. We hypothesize that all interesting concepts are individual candidates for future subtopics in the taxonomy.

Definition Let C^t be a set of interesting concepts obtained at time point t. Then, for all $C_i^t \in C^t$, C_i^t is reported to be a new topic in $T^{t'}$ where $t < t'$.

In Figure 2, the concept $C = \{$*nuclear, disaster, Fukushima, Japan, Daiichi*$\}$ is identified as a new topic.

3.2 Topic-based Prediction

This type of prediction is based on the hypothesis that interesting concepts do not necessarily represent new topics by themselves, but rather identify the taxonomic contexts where new topics should appear. To this end, we mapinteresting concepts to their best matching topics in the current taxonomy version and predict that these topics will exhibit sub-topics in future taxonomy versions.

Definition Let C^t be a set of interesting concepts at time point t, and let T^t be the current taxonomy version. A topic $k_j \in N(T^t)$ is *hot*,if there is $c^t \in C^t$ such that more than a fraction of τ_{match} of the c^t terms is contained in $Ext(k_j)$.

Definition Let K be a set of hot topics in T^t. Then, for all $k_i \in K$, k_i is reported as a topic which will undergo the creation of a new concept change in $T^{t'}$, where $t < t'$.

In the example shown on Figure 2, the topic-based predictor determines that the concept *"nuclear, disaster, Fukushima, Japan, Daiichi"* matches two DMOZ topics, *"Top/Science/Safety and Accidents"* and *"Top/Science/Nuclear Energy"*. These are the currently "hot" topics, which are likely to obtain a new child topic in future versions of DMOZ. However, the actual change may as well be more coarsely located in *"Top/Science/"*.

4. EXPERIMENTAL EVALUATION

4.1 Datasets and System Configuration

All components of the PIWO system are implemented in Java 1.6.0 using Cloudera's distribution CDH3 of Hadoop 0.20.2. For our experiments we used the New York Times (NYT) archive for concept finding. This archive contains about 1.5 million daily articles, spanning 20 years from 1^{st} *January* 1987 to 9^{th} *June* 2007. The total size of the vocabulary is ca. 1 million terms. For the prediction task we used the snapshots of the Open Directory Project `dmoz.org`. There are ca. 80 DMOZ snapshots between *February* 2003 and *August* 2009. For a meaningful prediction, we focused on the time overlap between the NYT archive and the DMOZ snapshots on a weekly basis, covering 232 weeks between January 2003 and June 2007.

4.2 Experimental Setup

We ran the pairwise term-relatedness computation algorithm. All term pairs with cosine similarity higher than $\tau_{rel} = 0.9$ were considered as related. We traced the mined candidates over a sliding window of six weeks. The deviation threshold value τ_{dev} was set to 0.4. We considered concepts existing for at least two weeks ($\tau_{cogn} = 2$). So we accepted concepts only if they appeared at least twice a week on average.

We considered all DMOZ subtopics related to health ("Health" branch), business ("Business" branch), and technology ("Science/Technology" branch). To define the ground-truth for the new-concept prediction, we identified the set of topics which did not exist in the previous snapshot of DMOZ. We

Space/Missions/Unmanned/Earth Observing/Aqua
Space/Missions/Unmanned/Earth Observing/CALIPSO
Space/Missions/Unmanned/Earth Observing/CloudSAT
Space/Missions/Unmanned/Earth Observing/Glory
Structural Engineering/Bridge/Failures/Minneapolis

Table 3: An example of sub-categories added in the *Top/Science/Technology/* branch.

Space
Space/Spacecraft and Satellite Design

Table 4: An example of topic-based predictor output for the *Top/Science/Technology/* branch.

performed some data cleaning by excluding all empty topics and auxiliary topics (not populated with links). In addition, we paid attention to the fact that a large portion of newly added topics is merely caused by renaming, moving, merging, or splitting operations. Such topics inherit the links from the previous version. To exclude these types of changes, we selected only new topics which contained new Web links not present at all in the previous snapshot of DMOZ.

[14] showed that $\mathcal{O}(3^{n/3})$ is the upper bound on maximal cliques in a graph with n nodes. In our computations, the number of available reducers r lessens this complexity by factor r. The overhead caused during the MapReduce job is $\mathcal{O}(f(n^3))$ for lexicon size n and Hadoop-dependent run time cost function f, taking into account sorting and splitting the input data.

4.3 Results

We experimented with two prediction strategies: concept-based and topic-based. We considered the use case of aiding Web catalogue editors. Therefore, we aimed at achieving high precision; recall was not a priority. NYT is mainly dedicated to US news with emphasis on the financial and business domains. Therefore, high coverage of all fine-grained topics (e.g., under health or technology) is infeasible.

Concept-based prediction considers all mined concepts as individual topics to be added to DMOZ. Table 1 shows the precision and recall of such concepts being really added at different levels of the taxonomy. We achieved pretty good precision between 67% and 98%. For example, PIWO correctly predicted a new concept *a-h1n1* in spring 2003. This is semantically very close to *SARS*, which was indeed added to DMOZ around this time.

Topic-based prediction aims to predict the existing topic where a new concept will be added. Table 2 shows the results. At high levels in the taxonomy, PIWO achieves high precision between 73% and 95%. For example (cf. Table 3 and Table 4), we correctly predicted the new sub-topic on *a-h1n1* under */Top/Health/ Conditions and Diseases/*. Another example is the emerging concept {*launching, space, international*} that caused PIWO to predict changes in the *Top/Science/Technology/Space/* topic. Indeed, new sub-topics *CALIPSO* and *CloudSAT* were added two months later, corresponding to launched satellites CloudSAT and international CALIPSO.

5. RELATED WORK

Topic Modeling. Latent topic models [3], like LSI [6, 12] or LDA [2], capture the joint distribution of terms (or other observable features) and documents such as Web pages. Clustering methods and matrix-factorization techniques also fall into this wider class of models. They reduce the dimensionality of the underlying co-occurrence data and thus bring out the most important concepts in unlabeled form. However, there is no topical hierarchy and there is no consideration of the data's dynamics. The ThemeMonitor of [18] explores topic evolution in document collections. A set of disjoint clusters is constructed for each collection snapshot and traced over time. Long-living clusters are interpreted as topics.

All of the above models require predefining the number of latent concepts (clusters, topics) – quite a limitation when dealing with real-life data at large scale. Our model does not make any assumptions of this kind.

Dynamics Modeling. There are numerous time-series-based models of data dynamics. In particular, there is ample work on detecting general or topic-specific bursts in social media. [11] models the temporal behavior of a topic using an infinite-state automaton. Each state corresponds to the particular intensity of topic mentions. Then a burst is a sequence of high-intensity states. Based on this seminal work, [21] discovers bursts of specific tag co-occurrences in collaborative tagging systems. Such bursty events are mapped to a taxonomy based on a tag hierarchy.

[17] computes correlations between micro-blogging activitiesand stock market events. [13, 1] develop methods for detecting emergent topics in blogs and tweets. Their notions of popular or emergent tag sets are related to our approach for mining interesting concepts. However, the focus of this work is on the real-time efficiency at the expense of using relatively simple models for emerging topics. Our model is more sophisticated by considering variable-cardinality and overlapping term sets, with additional considerations on temporal stability and saliency.

[15] study features of Twitter messages associated with a trend. We analyze the temporal behavior of concepts, but not the features of documents in the collection.

None of the above work considers pre-specified taxonomies like those of Web catalogues.

Prediction of Structural Changes in Taxonomies. Predicting topic additions to DMOZ has been considered in [4]. The approach is based on the hypothesis that a new subtopic is created when its parent topic contains multiple groups of tightly related Web links, leading to a topic split. The prediction is solely based on the state of the taxonomy itself. In contrast, we consider information that is external to the taxonomy, such as news, and predict totally new topics.

[10] develops methods for automatically organizing a stream of incoming documents into a taxonomy, and interprets new documents that do not fit any category as a trigger for creating a new topic.Our approach treats the DMOZ structure as a gold standard, and proposes new topics only if there is substantial evidence that the current categories are insufficient.

[20] automatically builds a probabilistic taxonomy from Web contents. This is carried out as a single batch computation. There is no consideration to the dynamics of the taxonomy. [5] builds an evolving taxonomy from social tags of Web pages, using techniques from association rule mining. [7] builds timelines of tag clouds from a stream of tagged pages, but does not impose any taxonomic structure on them.

Top/Health			Top/Business			Top/Science/Technology		
Depth	Precision	Recall	Depth	Precision	Recall	Depth	Precision	Recall
3	0.857	0.214	3	0.669	0.249	3	0.976	0.488
4	0.625	0.048	4	0.438	0.0975	4	0.500	0.138
5	0.333	0.009	5	0.266	0.040	5	0.229	0.087

Table 1: Concept-based prediction results

Top/Health			Top/Business			Top/Science/Technology		
Depth	Precision	Recall	Depth	Precision	Recall	Depth	Precision	Recall
3	0.847	0.324	3	0.730	0.239	3	0.952	0.476
4	0.695	0.136	4	0.462	0.0953	4	0.502	0.126
5	0.353	0.028	5	0.218	0.033	5	0.181	0.073

Table 2: Topic-based prediction results

6. CONCLUSION

The evolution of Web catalogues – representing the collective memories of society on the Web – can be predicted by detecting emerging latent concepts in the news. The PIWO system automatically discovers such concepts predicts new topics to be added to the Web catalogue's taxonomy. Our experiments showed that PIWO achieves high precision when predicting changes for the third and fourth taxonomy levels. The proposed mining and prediction methods can be applied to aid Web catalogue editors and for automatic taxonomy extension.

Acknowledgements

This work is supported by the 7^{th} Framework IST programme of the European Union through the focused research project (STREP) on Longitudinal Analytics of Web Archive data (LAWA) under contract no. 258105.

7. REFERENCES

[1] Foteini Alvanaki, Sebastian Michel, Krithi Ramamritham, Gerhard Weikum: EnBlogue: emergent topic detection in web 2.0 streams. SIGMOD 2011

[2] David M. Blei, Andrew Y. Ng, Michael I. Jordan: Latent Dirichlet Allocation. Journal of Machine Learning Research 3: 993-1022, 2003

[3] David M. Blei. Introduction to probabilistic topic models. Communications of the ACM, 2012

[4] Janez Brank, Marko Grobelnik, and Dunja Mladenić. Predicting category additions in a topic hierarchy. ASWC 2008

[5] Bin Cui, Junjie Yao, Gao Cong, and Yuxin Huang. Evolutionary taxonomy construction from dynamic tag space. WISE 2010

[6] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, Richard A. Harshman: Indexing by Latent Semantic Analysis. JASIS 41(6): 391-407 (1990)

[7] Micah Dubinko, Ravi Kumar, Joseph Magnani, Jasmine Novak, Prabhakar Raghavan, Andrew Tomkins: Visualizing tags over time. TWEB 1(2), 2007

[8] Robert D. Edwards and John Magee. *Technical analysis of stock trends*. J. Magee, 1997.

[9] Dan He and D. Stott Parker. Topic dynamics: an alternative model of bursts in streams of topics. KDD 2010

[10] Han-joon Kim and Sang-goo Lee. An intelligent information system for organizing online text documents. *Knowledge and Information Systems*, 6:125–149, 2004

[11] Jon Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7:373–397, 2003

[12] Thomas K. Landauer, Danielle S. McNamara, Simon Dennis, and Walter Kintsch. *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates, 2007

[13] Michael Mathioudakis, Nick Koudas: TwitterMonitor: trend detection over the twitter stream. SIGMOD Conference 2010: 1155-1158

[14] John W. Moon and Leo Moser. On cliques in graphs. *Israel Journal of Mathematics*, 3:23–28, 1965

[15] Mor Naaman, Hila Becker and Luis Gravano Hip and trendy: Characterizing emerging trends on Twitter. *Journal of the American Society for Information Science and Technology* , 62(5):902–918, 2011

[16] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[17] Eduardo J. Ruiz, Vagelis Hristidis, Carlos Castillo, Aristides Gionis, and Alejandro Jaimes. Correlating financial time series with micro-blogging activity. WSDM 2012

[18] Rene Schult and Myra Spiliopoulou. Discovering emerging topics in unlabelled text collections. ADBIS 2006.

[19] Bin Wu, Shengqi Yang, Haizhou Zhao, and Bai Wang. A distributed algorithm to enumerate all maximal cliques in mapreduce. In *Frontier of Computer Science and Technology, 2009*

[20] Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q. Zhu. Probase: A probabilistic taxonomy for text understanding. Technical Report, Microsoft Research Beijing, 2011

[21] Junjie Yao, Bin Cui, Yuxin Huang, and Yanhong Zhou. Bursty event detection from collaborative tags. *World Wide Web* 15(2): 171–195, 2012

[22] Yunyue Zhu and Dennis Shasha. Efficient elastic burst detection in data streams. KDD 2003