Evolution of Semantically Identified Topics

Victor Mireles and Artem Revenko

Semantic Web Company, Vienna Austria,
{victor.mireles-chavez,artem.revenko}@semantic-web.com,

Abstract. Topics in a corpus evolve in time. Describing the way this evolution occurs helps us to understand the change in the prominence of concepts: one can gain intuition about what concepts become more important in a given topic, which substitute others, or which concepts become related. By defining topics as weighted collections of concepts from a fixed taxonomy, it is possible to know if said evolution occurs within a branch of taxonomy or if hitherto unknown semantic relationships are formed / dissolved with time. In this work, we analyze a corpus of financial news and reveal the evolution of topics composed of concepts from the STW thesaurus. We show that using a thesaurus for building representations of documents is useful. Furthermore, the different abstraction levels encoded in a taxonomy are helpful for detecting different types of topics.

Keywords: topic discovery, taxonomy, thesaurus, topic evolution, topic modelling

1 Introduction

Of the many dimensions that can be ascribed to text corpora, the topics that they deal with is a very intuitive one for human readers. In a sense, topics constitute subgraphs of a "platonic knowledge graph", where only certain concepts and certain relations exist. Two qualities of topics are of particular interest: they reflect the intentions and context of the author of a text, and they are often treated in different texts by different authors. For these reasons, understanding their evolution in time can be treated as a proxy for studying the context of the authors.

In the case of corpora of news articles, understanding the evolution of topics can give insights into which entities become important in a given topic, or how they loose their importance. Furthermore, the detection of emerging and fading topics can be of interest for signalling major events.

In this work, we approach the study of topic evolution by using controlled, semantically enriched vocabularies. With our method, it is possible to describe the topics present in a certain time in terms of the topics present in the previous time points. With this in hand, the method is able to recover stable topics that are consistently dealt with in the news. Furthermore, detection of important events that shift the composition of topics is also possible. Finally, we perform an analysis of the topic-identification power of different levels of abstraction, as defined by a thesaurus.

1.1 Topic Discovery

Topic discovery, also known as topic modelling, is the task of analyzing a corpus and extracting from it clusters of terms that are semantically related.

When approached with statistical tools, the semantic relationship of the discovered topics is deduced from their distributional properties: terms that co-occur more often are deemed to be semantically related. Usually topic discovery is approached by first representing documents in terms of bags of words, n-grams[16] or embedded representations[11], forming from such representation a document-term matrix and, finally, inferring topics from said matrix. This last step is performed by matrix decomposition methods such as SVD (a.k.a. LSA[9]) or NMF[13], or by generative probabilistic models such as LDA[2] or PLSA[8]. The outcome of topic discovery is a collection of sets of terms, called topics, such that each document in the corpus can be assigned, often with a certain probability, to one of the topics.

In the above described scenario, the only semantic relations between the terms that we have access to are those statistically discovered based on the document-term matrix. However, in many applications further semantic relations are known between the terms. For example, if information about synonyms is known, topic detection can be done by counting occurrences of synsets[6].

In this work, we aim at incorporating further semantic information into the topic discovery process by use of a thesaurus. A thesaurus is a controlled vocabulary whose concepts are organized according to their hypernym/hyponym relations. In effect, it is a multihierarchical directed graph, where a node represents a concept and edges represent hypernym or hyponym relations. Each concept is assigned one or more labels: strings that can be matched against a document.

1.2 Topic Evolution

Studying topic evolution can be seen as a *study of the history of ideas*[7]. By performing topic discovery on several corpora, each of which has a timestamp, it is possible to see the transitions in interests in the author(s) of the corpora. This might be useful to discover how a given topic is treated differently in different times, thus constituting a proxy to study the evolution of semiotics. Several approaches have been adopted to the study of topic evolution. Some perform topic discovery independently in every corpus and only afterwards analyze the relationships between them (e.g. [7]) while others perform topic discovery in a corpus based on the topics discovered in the previous one (e.g. [1, 15, 14]). The former approach is subject to the variation inherent to topic discovery methods, which, in particular, can lead to topics being ”lost” from one corpus to the next due to corpus quality or size. This can lead to the independently discovered topics being difficult to compare. The latter approach has two main limitations: 1) that new ”flash topics” are hard to detect, and 2) they become over-sensible to parameters, such as thresholds for estimating the number of topics. However, in the preliminary experiments we have confirmed that dynamic topic models and plain NMF with subsequent estimation of transition between different time points yield similar results.

In this work, we present an intermediate approach. Topics are discovered independently for each corpus, and corpora in successive times are analyzed to determine in which ways did the topics transit into others or appeared de-novo. This second step allows us to describe the evolution of topics not just as the evolution of sets of co-occurring terms, but rather as the merging and splitting of existing topics. Hence, we are enabled to define the notion of *persistent* topic, i.e. topics that appear in several consecutive corpora.

2 Data

The dataset we analyze is a financial news data set. The news come from a single source (Bloomberg news) and are made available as supporting data in[4]. From the original dataset, we took articles between January 2009 and November 2013, which total 447,145 documents. We consider the documents within each week, starting on Friday, a different corpus. Only a subset of the original dataset was used, in order to guarantee that all corpora have at least 50 documents. The sizes of corpora range between 50 and 4900, with a mean corpus size of 2589 documents.

The semantic relationships between concepts that we are using in this work are those expressed by *skos:narrower* and skos:broader predicates in version 9.02 of the STW Thesaurus for Economics[3]. This thesaurus consists of 6221 concepts, of which 4108 are leaves (i.e. they have no narrower concepts). The wide range of concepts included in the thesaurus make it ideal for analyzing the corpus of financial news, specially because of its concept schemes of Geographic Names and General Descriptors. For the purposes of this work, the predicate skos:topConceptOf is considered to be equivalent to skos:broader.

3 Methods

Entity extraction Entity extraction was performed using PoolParty Semantic Suite[1]. In brief, the texts are pre-processed in the same manner as the labels

[1] `poolparty.biz`

from the thesaurus, namely, stopwords are removed, tokens are lemmatized, n-grams up to 4-grams are constituted. Then, a matching is done to identify all the concepts appearing in the documents. Thus, for every concept c in the thesaurus and every document, we have computed $n_d(c)$, the number of times any label of c appears in document d. Finally, all the documents corresponding to the same week were put together into a single corpus, which we will denote by $\mathcal{C}_w$, where w is the week number.

Representing documents For each document d in each corpus, two vector representations are computed.

The first, which we call *level 0* representation, contains information only about a subset of all concepts that have no narrower concept in the thesaurus. We call this set the set of leaves, and denote it by $l_1, l_2, ..., , l_{m_0}$. The level 0 representation of a document d is then a m_0-dimensional vector $V_0(d)$ whose i'th entry is given by $V_0(d)[i] = \frac{n_d(l_i)}{n_d}$, where n_d is the number of tokens in document d.

The second representation, called *level 1* representation, contains only information of those concepts that are broader concepts of some leaf. If we denote that set by $b_1, b_2, ..., b_{m_1}$, then this representation consists of a m_1 dimensional vector, whose i'th entry is given by

$$V_1(d)[i] = \sum_{l \in L(b_i)} \frac{n_d(l)}{n_d} \tag{1}$$

where $L(c)$ denotes the set of nodes which are narrower than concept c.

It must be noted that with both of these representations only the occurrences of leaf concepts is considered. The difference being that in level 1 representations the occurrence of concepts that are narrower of one same concept are grouped together. With these two representations defined, we can represent each corpus C_w by two matrices: $A_0(w)$ and $A_1(w)$, both of which have as many columns as documents in corpus $\mathcal{C}_w$, but with m_0 and m_1 rows respectively.

Detecting topics in a week For each week w, we compute a Non-Negative Matrix Factorization (NMF)[10] on the two matrices $A_0(w)$ and $A_1(w)$. NMF decomposes a matrix $A \in \mathbb{R}_+^{m \times n}$ into the product TS of two matrices, with $T \in \mathbb{R}_+^{m \times k}$ and $S \in \mathbb{R}_+^{k \times n}$. To choose the value of k, we first compute $\sigma_1, \sigma_2, ..., \sigma_n$ all singular values of A in descending order, and then choose the k that maximizes $\frac{\sigma_k - \sigma_{k+1}}{\sigma_{k+1} - \sigma_{k+2}}$. This method is equivalent to the *eigenvalue gap trick*[5] in the case of non-negative matrices. The NMF decompositions were computed using the scikit-learn library[12] setting the parameter *l1ratio* to 1. We followed NMF by a sparsifying step: the smallest threshold θ was found (using gradient descent) such that if the matrix T_θ is the result of setting all entries of T whose value is lower than θ to 0, then the density of $T_\theta S$ is not more than one half the original density of A. From now on, we refer to T_θ simply as T.

The use of NMF yields, for each week w two pairs of matrices: $T_0(w), S_0(w)$, and $T_1(w), S_1(w)$. We call the matrices T resulting from NMF the *concept-topic matrices*. If $T[c, j] > 0$, we say that concept c belongs to topic j with a degree of $T[c, j]$. Thus, for every week w we can compute two sets of topics, one for each representation.

Detection of Topic Transitions For a representation $q \in \{0, 1\}$ and two consecutive weeks we get from the previous steps the matrices $T_q(w) \in \mathbb{R}_+^{m_q \times k_1}$ and $T_q(w{+}1) \in \mathbb{R}_+^{m_q \times k_2}$, where k_1 and k_2 are the numbers of topics in different weeks. In order to detect transitions between topics in these two consecutive weeks, we solve the optimization problem of finding the matrix $M_q(w) \in [0, 1]^{k_1 \times k_2}$ that minimizes $||T_q(w)M_q(w) - T_q(w + 1)||_2$. The resulting matrix $M_q(w)$, which we call *transition matrix*, expresses each topic in week $w + 1$ as a linear combination of the topics in week w.

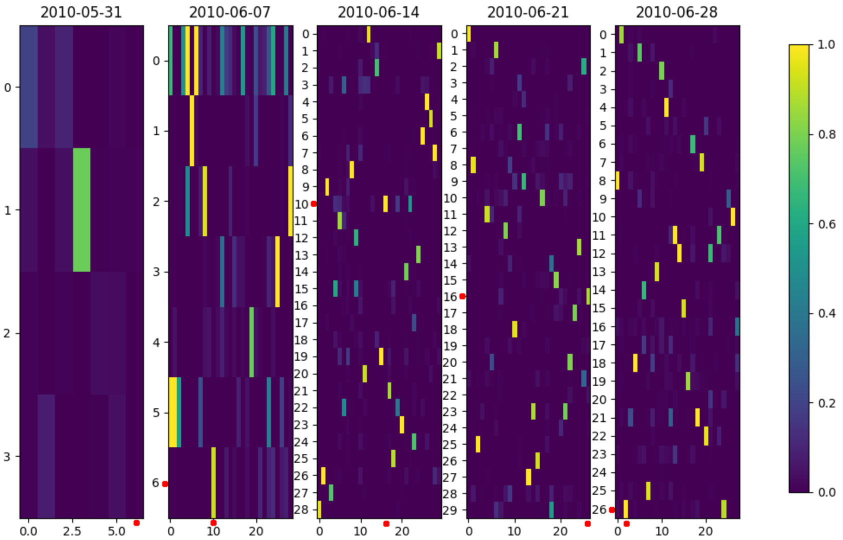

Fig. 1. Transition matrices between 6 consecutive weeks. These are the weeks in which the Football World Cup flash topic is detected. The topic is not present in the first week, and its index within each of the following weeks is 6, 10, 16, 26 and 2 (indicated with red dots)

The following points help in the interpretation of transition matrices:

- If two topics merge into a new topic in the next week, then the new topics column will have large entries in the rows corresponding to the two previous topics.

- A new topic in week $w+1$ will correspond to a column whose entries are all small: it is not similar to any topic in the previous week
- A topic whose concepts don't change between two consecutive weeks can be detected by a column and row that both have a single entry close to 1.
- One can think of topic transition as the process of the topic from the previous week distributing its weight into the topics of the current week. The topic cannot give more weight than it has.

With a set of consecutive transition matrices $M_q(w), M_q(w+1), \ldots, M_q(w+g)$, it is possible to detect topics which remain stable for several weeks: they will be a sequence of indices $t_1, t_2, ..., t_g)$ such that $1 - M_q(w)[t_i, t_{i+1}] \leq \alpha$ for $i = 1...g-1$ and some small α. We consider a topic to be a *stable topic* if the above condition holds for $g >= 5$ with $\alpha = 0.2$, i.e. if the topic keeps approximately with the same concept composition for at least 4 weeks. Figure 1 is an example of a set of transitions matrices that exhibit a stable topic.

4 Results

We decomposed all corpora into topics leading to different number of topics per week (Fig 2). After computing the corresponding transition matrices we are able to detect the appearance of new topics as well as several stable topics. Among them, we can talk of *persistent* and *flash* topics.

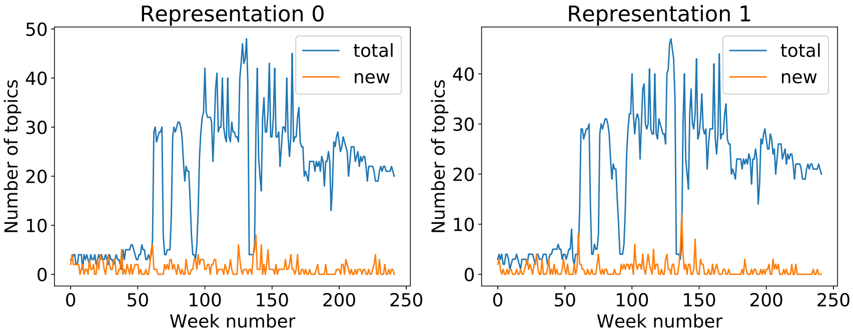

Fig. 2. Number of topics discovered for every week. The dips are due mostly to weeks in which few documents were collected. A topic is considered new if it belongs to no earlier stable topic.

Persistent topics are topics that the news source treats regularly. While interruptions in the data (weeks with few articles) sometimes fragmented these topics, in the sense of our definition of stable topics, they were quick to recover. Several stable topics represented through their most important concepts are presented in Table 1.

Table 1. Twelve most important concepts in Persistent topics. Concept are sorted in decreasing order of average (per week) importance. Interestingly, *Bayesian Inference* appears as part of a topic, because *Prior* is an skos:altlabel for this concept in the thesaurus.

Stocks	Futures	Japanese	Euro	Meat
Market	Future	Yen	Germans	Beef
Stock market	Soybean	Tokyo	German	Light
Product	Crops	Japanese	European	Bayesian inference
International	Gold	Loss	Greek	Cattle
Purchase	Wheat	Electronics	Greeks	Price
Market value	Department	Newspaper	Greece	Department
Price	Sugar	Sales	Berlin	Plants
Swap	Rubber	Dividend	Nation	Flavour
Benchmarking	Singapore	Services	Bailout	Sales
Hedging	Cocoa	Product	Economy	Product
Loss	Cattle	Semiconductor	Portuguese	Import
Future	Palms	Plants	London	Tokyo

Furthermore, we were able to detect Flash topics, that is topics which relate to specific, transient events. We found the following flash topics in level 0 representation:

1. the 2010 South Africa World Football Cup,
2. the 2010 artillery fire exchange in the Korean peninsula.
3. the 2011 Drought,
4. the Arab Spring,
5. the Fukushima Daiichi nuclear disaster,

The most frequently mentioned concepts in these topics can be seen in table 2.

The evolution of topics can best be exemplified by the Football World Cup example. In table 3, the change in the topic across weeks is shown. Notice how the number of countries decreases, and those which remain are also those which remained in the tournament. Let us recall that the tournament ended on the 11th of July.

Many of these topics were found also in the level 1 representation. In general, the level 1 representation yields longer stable topics, as can be seen on the length distributions of stable topics in Figure 3. It is worth mentioning, that a topic which is very fragmented (in time) in the level 0 representation is less so in the level 1 representation: that of the Japanese market. This suggests that looking at more abstract concepts can increase stability of topic detection. It is important that, while using the broader concept of a set of leaves increases stability, there is no evidence that similar topics are confounded by this process. This is an indication that topics are not necessarily matching branches of the taxonomy but, rather, are combinations of topics from across the thesaurus.

Table 2. Twelve most important concepts in Flash topics. Concepts are sorted in decreasing order of average (per week) importance.

Football 6 Weeks	Korea 5 Weeks	Drought 5 Weeks	Arab Spring 8 Weeks	Fukushima 6 Weeks
World	Koreans	Wheat	Libya	Plants
Sport event	Korean	Crops	International	Nuclear energy
South Africa	South Korea	Drought	Nation	Manufacturing plant
Football	North Korea	Soybean	Foreign	Electricity
African	South Korean	World	Arabs	Cooling
Matching	South Koreans	Rice	Arab	Earthquake
Coaching	Officials	Food price	Egypt	Greenhouse gas emis.
South African	Nation	Nation	Air	Nuclear power plant
South Africans	Island	Egypt	West Asia	Nuclear fuel
Brazil	Foreign	Department	Tunisia	Order
Argentina	World	Australia	Officials	Process
Netherlands	Chinese	International	African	Officials
Mexico	Fire	Province	Industrial action	Fire

For the same reason, topics in consecutive weeks which are not deemed persistent under the level 0 representation, become so in the level 1 representation. It is thus important to choose carefully the level of granularity of the concepts that will be used to annotate a corpus with the aim of persistent topic detection. It must also be noted that detecting topics based solely on the words (i.e. without a controlled vocabulary) does not provide this possibility.

Finally preliminary results show that, on average, concepts gained by a topic during its evolution are closer in the thesaurus to already existing topics that is expected at random.

5 Conclusions and Future Work

We have presented a method that is able to detect both persistent and transient topics in news sources. Interestingly, we have shown that it is possible to detect such topics both when annotating documents only with leaves from a thesaurus, and when annotating them also with those concepts directly above leaves. Furthermore, we have shown that the relatively simple NMF method is able to detect stable topics, and that this stability can also be captured by our proposed method for computing topic transitions. By considering the news articles in each week as independent corpora, we are able to detect short-lived topics, that would otherwise be lost in a global topic modeling.

The fact that topics are detectable in both representations is an indication that topics are not easily confounded with each other when one considers more abstract categories. This is an interesting result, for it shows that the concepts comprising a topic are distributed widely enough across the thesaurus that abstracting them just one level still allows for their detection. In a sense, the way that the concepts have been organized in the thesaurus do not match the real-world occurring topics. We believe that this result can serve as a measure of the

Table 3. Evolution of the concepts in the Football World Cup Topic. The start date of each corpus is shown in the header. All concepts belonging to the topic are shown, sorted in decreasing importance.

2010-06-07	2010-06-14	2010-06-21	2010-06-28	2010-07-05	2010-07-12
World	World	World	World	World	South Africa
Sport event	Sport event	Sport event	Sport event	South Africa	African
Football	Matching	Football	Football	Sport event	World
Matching	Football	South Africa	Coaching	Football	South African
South Africa	South Africa	Coaching	South Africa	African	South Africans
Slovenia	Brazil	Mexico	Matching	South African	Platinum
Italy	Coaching	France	Brazil	South Africans	Football
Algeria	Argentina	Brazil	Argentina	Humans	Sport event
Australia	Mexico	Argentina	Ghana	Uganda	British
Paraguay	France	African	Uruguay	Black people	Iron
Netherlands	French	Uruguay	Netherlands	Somalia	Iron ore
Ghana	Slovenia	Ghana	Spain	Matching	International
Nation	Cote d'Ivoire	Netherlands	Punishment	International	Golf
Cameroon	Algeria	French	Paraguay	American	AIDS
Coaching	Portugal	Italy	African		Black people
African	African	Loss	Loss		
Brazil	Uruguay	American	Nigeria		
London	Serbia	Slovenia	Gambling		
Serbia	Italy	Slovakia	Federation		
Industrial action	New Zealand	Chile	Nation		
South Korea	Nigeria		Industrial action		
Punishment	South African		American		
North Korea	South Africans		International		
Economy	Slovakia				
Greece	Punishment				
Spain	Greece				
South African	American				
South Africans	Chile				
Police	Netherlands				
Islamic					
Dutch					
New Zealand					
Slovakia					
Argentina					
Mexico					
Portugal					

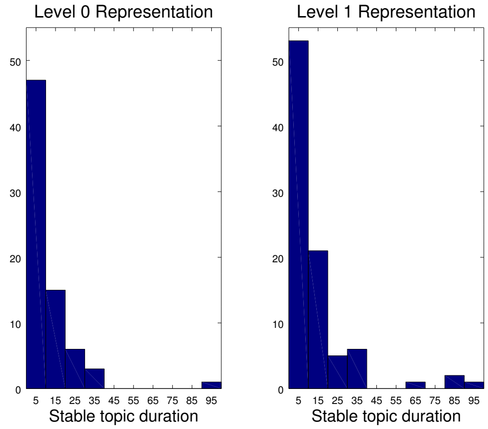

Fig. 3. Distribution of the lengths of stable topics. Level 1 representation shows both more and more long lasting stable topics. Considering the broader of a leaf concept thus reduces noise in the description of documents.

generality of a thesaurus, and its applicability to analyzing texts from various sources.

This work represents initial results in the analysis of topic transitions with the help of a thesaurus. In future work we intend to build on top of the current results and extend the methodology. In particular, the current results motivate us to:

1. Investigate and compare representation of topics in different levels. The preliminary observations suggest that several stable topics at level 0, in non-overlapping weeks, could be merged into a single topic at level 1.
2. The topics at level 1 appear to be more stable. This could be useful in the case of limited data, when the detailed topic could fade away.
3. The different levels of representation could prove to be useful for people with different backgrounds. Namely, we expect that more detailed topics could be of interest for experts in the field, whereas the more general topics could give a good overview for less experienced readers.
4. We aim at improving the transition computation with the help of taking the distances between concepts into account and employing methods similar to soft cosine similarity.
5. Finally, statistical analysis is required to confirm or observations that the concepts gained during topic evolution are more likely to be close, in the thesaurus, to concepts already in a topic.

Acknowledgements The work is partially supported by the PROFIT (`http://projectprofit.eu/`) project. Part of the European Commission's H2020 Framework Programme, grant agreement no. 687895,

References

1. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the 23rd international conference on Machine learning. pp. 113–120. ACM (2006)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3(Jan), 993–1022 (2003)
3. Borst, T., Neubert, J.: Case study: Publishing stw thesaurus for economics as linked open data. W3C Semantic Web Use Cases and Case Studies (2009)
4. Ding, X., Zhang, Y., Liu, T., Duan, J.: Using structured events to predict stock price movement: An empirical investigation. In: EMNLP. pp. 1415–1425 (2014)
5. Djurdjevac, N., Sarich, M., Schütte, C.: Estimating the eigenvalue error of markov state models. Multiscale Modeling & Simulation 10(1), 61–81 (2012)
6. Ferrugento, A., Alves, A., Oliveira, H.G., Rodrigues, F.: Towards the improvement of a topic model with semantic knowledge. In: Portuguese Conference on Artificial Intelligence. pp. 759–770. Springer (2015)
7. Hall, D., Jurafsky, D., Manning, C.D.: Studying the history of ideas using topic models. In: Proceedings of the conference on empirical methods in natural language processing. pp. 363–371. Association for Computational Linguistics (2008)
8. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine learning 42(1), 177–196 (2001)
9. Landauer, T.K., Laham, D., Foltz, P.W.: Learning human-like knowledge by singular value decomposition: A progress report. In: Advances in neural information processing systems. pp. 45–51 (1998)
10. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401(6755), 788 (1999)
11. Nguyen, D.Q., Billingsley, R., Du, L., Johnson, M.: Improving topic models with latent feature word representations. Transactions of the Association for Computational Linguistics 3, 299–313 (2015)
12. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
13. Recht, B., Re, C., Tropp, J., Bittorf, V.: Factoring nonnegative matrices with linear programs. In: Advances in Neural Information Processing Systems. pp. 1214–1222 (2012)
14. Saha, A., Sindhwani, V.: Learning evolving and emerging topics in social media: A dynamic nmf approach with temporal regularization. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining. pp. 693–702. WSDM '12, ACM, New York, NY, USA (2012)
15. Vaca, C.K., Mantrach, A., Jaimes, A., Saerens, M.: A time-based collective factorization for topic discovery and monitoring in news. In: Proceedings of the 23rd International Conference on World Wide Web. pp. 527–538. WWW '14, ACM, New York, NY, USA (2014)
16. Wang, X., McCallum, A., Wei, X.: Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In: Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on. pp. 697–702. IEEE (2007)