Towards an Evaluation Framework for Topic Extraction Systems for Online Reputation Management[*]

Enrique Amigó[1], Damiano Spina[1], Bernardino Beotas[2], and Julio Gonzalo[1]

[1] Departamento de Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia
C/Juan de Rosal, 16
28020 Madrid, España
{enrique,damiano,julio}@lsi.uned.es
[2] Grupo ALMA
C/Valentín Beato, 23
28037 Madrid, España
b.beotas@almatech.es

Abstract. This work present a novel evaluation framework for topic extraction over user generated contents. The motivation of this work is the development of systems that monitor the evolution of opinionated topics around a certain entity (a person, company or product) in the Web. Currently, due to the effort that would be required to develop a gold standard, topic extraction systems are evaluated qualitatively over cases of study or by means of intrinsic evaluation metrics that can not be applied across heterogeneous systems. We propose evaluation metrics based on available document metadata (link structure and time stamps) which do not require manual annotation of the test corpus. Our preliminary experiments show that these metrics are sensitive to the number of iterations in LDA-based topic extraction algorithms, which is an indication of the consistency of the metrics.

1 Introduction

The growing interest on monitoring opinions in the Web 2.0 is well known. Online Reputation Management consists of monitoring the opinion of Web users on people, companies or products, and it is already a fundamental tool in corporate communication. A particularly relevant problem is to detect new topics or opinion trends which deserve the attention of communication experts, such as a burst of tweets or blog entries about a controversial issue about a company, or a defect of a product. A system that assists a communication expert should be able to detect (particularly new) topics, tag them in an interpretable way, cluster documents related to each topic and analyze the evolution of topics over time. What makes this a distinctive problem is the fact that documents are naturally multi-topic: relevance of a document for a topic may be even sub-sentential. This problem is sometimes referred to as Temporal Text Mining [1,2].

Models and systems to solve these tasks are recently starting to appear in scholar publications. But a major bottleneck so far is the absence of a benchmarking test suite to evaluate and compare systems. Creating such a gold standard is, in fact, a complex task: defining the set of topics in a document stream is a subtle task, because topics tend to co-occur in documents and the appropriate level of granularity in topic and sub-topic distinctions is something fuzzy to fix. For similar reasons it is also difficult, once the set of topics is established, to decide which documents talk about each of the topics and how central is each document to each of the topics that the document discusses. In the absence of a gold standard, extrinsic precision-recall based metrics can not be applied.

For this reason, current systems are evaluated informally via use cases, or otherwise using intrinsic evaluation measures which are specific to the model being tested.

There are, however, basic restrictions on how a good system should behave. For instance, documents which share outlinks to the same web pages should tend to be more related than documents which do not share outlinks. This type of information has not been yet used by current topic detection systems, because they relate together only a small subset of the documents. This information, however, might be used as a (limited) evaluation or validation mechanism to optimize system parameters. In this paper we address the task of defining an evaluation methodology based on this idea, and check its suitability on an LDA-based approach to topic detection over time.

2 State of the art

We will start with an overview of models to solve the task, and then we will summarize the evaluation methodologies used so far and discuss their limitations.

2.1 Topic detection models

The most basic approaches for topic monitoring focus on word frequency. The assumption is that frequent words indicate, in general, salient topics in a document

collection. Some available web services are Blogpulse Trends[3], Mood Views[4] and Blogscope[5]. Brooks and Montanez showed that frequent words (extracted according to tf.idf), produce tags that generate document clusters with more cohesion than user tags in blogs [3].

Gruhl included the temporal dimension in his model by extracting topic terms with frequency peaks over time [4]. Chi considered also the distribution of terms across blogs [5]. He assumed that topics gain prominence in blog subsets. His model consists of computing the singular values of the time-blog frequency matrix . Mei et al. combine topological information with the temporal dimension [6]. Their model employs the EM algorithm to identify the topic distributions along time and location that maximize the likelihood of word occurrences. An interesting feature of this model is that it assumes that several topics can appear in the same document.

Many novel proposals are currently based on the LDA (Latent Dirichlet Allocation) model [7]. As well as Mei's approach, LDA is a probabilistic model that estimates the distribution θ_d of topics for each document d and the distribution of words for each topic. The particularity of LDA is that the distribution parameters are generated by a Dirichlet distribution with certain hyperparameters that are stated a priori.

One example of these models is TOT (Topic Over Time) [8]. The most characteristic aspect of this work is that the temporal variable is added to the LDA model, assuming that topics follow a Beta distribution along time. One drawback in this work is that all the document collection must be processed for inferring temporal distribution when new documents appear in the input stream. The model *Dynamic Topic Model* [9] tries to solve it by estimating topic distributions for each time slot independently. After this, the model employs temporal series techniques in order to analyze topic evolution. Another model that tackle this issue is *On-line LDA* [10]. This model states that the knowledge produced over a time slot represents the a priori knowledge for the next time slot. This idea allows to process new documents without reprocessing the whole collection. However, an addition mechanism to detect new topics along the time becomes necessary. Another interesting model based on LDA is denominated *Multiscale Topic Tomography*[11]. In this approach the topic distribution includes different granularity levels.

2.2 Evaluation approaches

The main bottleneck in this area of research is the absence of a common evaluation methodology to compare approaches. Let us summarize the approaches to evaluation in the research described above.

In terms of its efficiency and suitability to assist experts in the online reputation management task, some approaches are best suited than others. For instance, the models *ON-line LDA* and *Dynamic Topic Model* are able to process

---

[3] www.blogpulse.com/trends

[4] ilps.science.uva.nl/MoodViews

[5] www.blogscope.net

new documents without re-processing the collection. *Multiscale Topic Tomography*, on the other hand, allows a topic visualization at different granularity levels. However, it is still necessary to define an evaluation framework to compare approaches in terms of accuracy.

Some approaches are simply evaluated over case studies. This is the case of Mei's approach [6] and the *Dynamic topic model* [9]. Ghuhl's model [4], on the other hand, is evaluated against human annotated topic terms; an evaluation method than can not, for instance, be applied to LDA-based models.

The model *Topic over Time* [8] is evaluated with intrinsic clustering metrics according to the KL-divergence between topics; but this methodology is only appropriate to compare similar systems. For instance, in their evaluation the authors obtained evidence about the advantages of including the temporal variable in the model. It is not possible, however, to evaluate heterogeneous systems with intrinsic clustering metrics. For instance, systems based on KL-divergence would be rewarded by this evaluation method. Something similar happens with the evaluation of the *Multiscale Topic Tomography*[11], where the perplexity of the model is compared against the perplexity obtained with other models. In this case, LDA-based models could not be compared with models based on traditional clustering algorithms. Other proposed evaluation metrics focus on extrinsic tasks using topic descriptors, such as multi-document summarization [2].

With these limitations in mind, our goal is to define and apply an automatic evaluation framework enabling comparison between heterogeneous, arbitrary systems, and which is not dependent on cost-intensive manual annotation of data.

3 Evaluation methodology

3.1 System prerequisites

We start from a few prerequisites for topic detection systems in opinion mining:

- **Aggregation:** The system must detect a finite number of topics. Documents will be associated to zero, one or more topics in a discrete or continuous way. The key point is that related documents should share at least one topic.
- **Temporality** In order to analyze the evolution of the reputation of a given entity, the system must reflect differences in topic distribution across time. This implies to show the intensity of topics across time slots.
- **Interpretability** Identified topics should be tagged in a way that is interpretable for the user.
- **Accessibility** For each topic, the corresponding documents must be ranked according to its relevance in the context of the topic.

In this work, we focus on the two first functionalities: "aggregation" and "Temporality". The interesting aspect of these two features is that it is possible to generate automatically a benchmark for evaluation purposes.

3.2 System Output variables

The aggregation functionality requires to infer to what extent each document is related to each topic. This output can be formalized as $P(\theta|d)$, which represents the distribution θ of topics in each document d. For instance, a traditional discrete clustering algorithm would return $P(\theta_i|d) = 1$ if the document d belongs to the cluster associated with the topic θ_i.

Analogously, the temporality function requires an output variable $P(\theta|t)$ representing the distribution of topics in each time slot t. From the perspective of evaluation, a key aspect is that all functionalities must be mutually consistent. In particular, the intensity of topics (temporality) has to correspond with the number of associated documents in the time slot (Aggregation). Therefore, temporality can be inferred from the output $P(\theta|d)$. Assuming that the intensity of topics is proportional to the number of related topic in the time slot, we can state that:

$$P(\theta|t) = \sum_{d \in t} P(\theta|d)$$

3.3 Evaluation measures

Our evaluation methodology is based on two assumptions on the desired behavior of systems:

- Documents with outlinks that point to the same page and documents produced by the same author will tend to be more topically-related than the average.
- It is easier to find highly related documents in the same time slot (say, blog posts in the same week), than separated by long time periods (such as several months).

As most current systems do not rely on this kind of information, it is possible to use it at least for parameter optimization cycles. Some of the systems do employ temporal information, and therefore the second restriction is not totally system-independent. In such cases, however, the improvement obtained by the use of temporal information can still be measured in terms of the first restriction.

The first step to evaluate systems according to these two assumptions consists of defining when the system considers that two documents are related (as for their topics). This is not straightforward, given that systems generate a distribution of weighted topics for each document. We will assume that one topic is enough to consider that two documents are related, but only if both documents focus on this topic. According to this, we define the *Connectedness* of a document pair as:

$$\text{Connectedness}(d1, d2) = \text{Max}_i(\text{Min}(P(\theta_i, d1), P(\theta_i, d2)))$$

Our evaluation metrics will compare the *connectedness* of document pairs in two sets according to the assumptions introduced above. We will call these sets RDP (Related Document Pairs) and NRDP (Non-Related Document Pairs).

RDP consists of document pairs with, for instance, one or more common out-links, while NRDP consists of documents, conversely, without common outlinks. According to our assumptions, document pairs in RDP should have a higher connectedness, in average, than document pairs in NRDP.

In order to avoid dependencies on scale properties of the distribution $P(\theta|d)$ associated to each system, we will formulate evaluation metrics in a non-parametric way, estimating, for each system s:

$$\text{metric value}(s) = P(\text{Connectedness}(d_r, d_r{'}) > \text{Connectedness}(d_n, d_n{'}))$$

where $< d_r, d_r > \in RDP, < d_n, d_n{'} > \in NRDP$.

In other words, the quality of the system is measured as the probability that two documents from the RDP set have a higher topic overlap (according to the system) than two documents from NRDP. Different criteria to form RDP and NRDP lead to different evaluation metrics; we now discuss some examples.

Outlink Aggregation In order to obtain the set of related document pairs (RDP), we assume that two documents are more likely to be related if they share an outlink to the same web page, if this outlink does not appear in other documents (this restriction eliminates frequent outlinks which are not related to the document content, such as links to Facebook).

Author Aggregation As for documents related by a common author, we will simply consider pairs of documents with the same author for RDP and pairs from different authors for NRDP.

Temporality As for temporality, we will assume that is easier to find documents sharing a topic when both documents belong to the same time slot. In particular, we will build RDP with the 100 most related document pairs (according to the system output) which are created in the same week. NRDP is formed by the 100 most related document pairs which are created with a difference of at least three months.

4 Test Case: Iterations in an LDA-based system

To test our evaluation methodology, we have implemented the LDA approach, starting with the algorithm described in [8] and eliminating the temporal variable component – which will be tested in future work –. LDA is a generative process where each document d is associated with a multinomial distribution of topics, and uses Dirichlet distributions as hyperparameters. The model assumes that each document token is associated to a single topic, and therefore the topic distribution in a document would be given by the individual token assignments. The article by Wang and McCallum describes the approach in detail as well as the derivation of the Gibbs sampling.

The algorithm implemented consists of the following steps:

1. Random initialization of each token to some of the k topics.
2. For each token in document d, the topic is updated drawing on the probability $P(z)$ for each topic z. The probabilities are:

$$P(z) = (m_{d,z} + \alpha)\frac{n_{z,w} + \beta}{\sum_v^V (n_{z,w} + \beta)}$$

   where $m_{d,z}$ represents the number of tokens in the document d associated to topic z; $n_{z,w}$ represents the number of occurrences of the word w from the corresponding token in topic z, and V is the vocabulary. α and β are two hyperparameters that reflect, respectively, the topic dispersion per word and per document.
3. $m_{d,z}$ and $n_{z,w}$ are updated and then we go back to step 2, for as many iterations as desired.

Implementations known to us use fixed hyperparameters for any word in the vocabulary and for every document. In future work, however, and counting with an automatic evaluation mechanism, we could test whether α should have some relation with the document length.

5 Hypothesis validation

In order to validate our assumptions, we have performed a small experiment which involves manual validation of document pairs.

From our testbed, we have generated 64 random tuples, each consisting of two document pairs: in one pair, both documents share at least one outlink that does not appear in any other document (see Section 3.3); in the other pair, they do not share any outlink. According to our hypothesis, document pairs which share outlinks should be more topically related in average than pairs that do not share outlinks.

For each tuple, we have manually annotated which is the most topically related document pair (sometimes this is not obvious and the tuple is then annotated as undecidable). For 50 tuples (78%), the document pair that share the outlink was more topically related than the other. In 12 cases (19%) it was undecidable, and only in 2 cases (3%) the linked document pair was less topically related than the other.

An analogous process was conducted for the co-authorship criterion, comparing tweet pairs written by the same author with pairs written by different authors. The results over 97 tuples are similar to the previous ones: co-authored tweets are more related in 80% of the cases, while non co-authored tweets are more related only in one occasion (1%).

These results suggest that our assumptions are reasonable for our testbed. This is, of course, just a preliminary result that must be validated with larger manual annotations over different testbeds. Note also that the experimental procedure must be refined, because "undecidable" cases (which are 20% of the assessed samples) might become decidable with a more precise, testbed-specific definition of relatedness.

6   Experiment and Evaluation Results

The goal of our experiment is to test the behavior of our evaluation metrics. As a dataset we use 5,000 tweets and 500 posts from blogs in Spanish containing the term BBVA (an Spanish bank operating in several countries). We have generated a vocabulary excluding stop words. In general, for all the approaches compared, topics detected by the LDA system consist of (i) information about "Liga BBVA", the Spanish Premier Football League, which is sponsored by the bank; (ii) economic information on the bank; (iii) information in languages other than Spanish (such as Catalan); and (iv) topics with unfrequent terms. In general, the granularity of the topics is relatively low. Note that, unlike other related experiments, we are focusing on a single entity, while other approaches cover several totally independent topics.

Table 1 displays the results of *Aggregation* (by outlinks) and *temporality* for LDA over 500 blog posts, fixing certain values of the hyperparameters and for different number of iterations. Note that aggregation goes from 0.41 up to 0.9, reaching a ceiling after 100 iterations. This number might be different when more documents are processed, or for different values of the hyperparameters. In any case, the results show a strong correlation between the number of iterations of the algorithm and the evaluation results; assuming that a higher number of iterations leads to better LDA results, our metric behaves consistently.

Table 1 also shows the *Temporality* obtained for different number of iterations. This metric also increases with the number of iterations, although this behavior is not so stable here. A possible reason is that the assumption that documents which are closer in time should be more related on average is not so valid as for the case of aggregation. Another possible reason is that this measure is estimated on the 100 most related document pairs only, while aggregation is computed on a larger set of samples.

Iterations	Agregation (by outlinks)	Temporality
1	0.41	0.42
5	0.60	0.35
10	0.76	0.6
20	0.86	0.57
50	0.89	0.60
100	0.90	0.61
200	0.90	0.61
500	0.90	0.65
1000	0.91	0.71
2000	0.90	0.72

Table 1. Evaluation results for 500 blog posts, 10 topics, $\alpha = 1$, $\beta = 0.1$ and different number of iterations

Table 2 shows the results on 5,000 tweets, this time measuring aggregation by author. Again, the metric values seem to stabilize around 100 iterations, and they show a clear correlation with the number of iterations.

Iterations	Agregation (by author)
1	0.47
5	0.55
10	0.61
20	0.72
50	0.76
100	0.78
200	0.79
500	0.8

Table 2. Evaluation results for 5000 tweets, 10 topics, $\alpha = 1$ and $\beta = 0.1$

Another variable that can be analyzed with our evaluation methodology is the effect of different values for the hyperparameter α. Table 3 shows that α does not have a strong effect on the results. In fact, the maximum seems to be around $\alpha = 1$. This implies that, in general, documents tend to be centered around one single topic. This is perhaps due to the low granularity of the topics produced in our experiment.

alpha value	Agregation (by outlinks)	temporality
0.1	0.89	0.64
1	0.9	0.66
5	0.9	0.67
10	0.89	0.66
20	0.88	0.74
50	0.84	0.62

Table 3. Evaluation results for 500 blog posts, 2000 iterations, 10 topics, $\beta = 0.1$ and different α values

Finally, we have studied the effect of the number of topics on the results of the evaluation. Is it possible that LDA, in this context, reaches a more adequate topic granularity by increasing their number? Table 4 shows the results obtained for 500 blog entries, 2000 iterations and $\alpha = 1$. Note that, although there is some positive effect when increasing the number of topics, it is not as clear as in previous experiments.

Number of topics	Agregation (by outlinks)	temporality
5	0.85	0.67
10	0.9	0.73
15	0.92	0.74
20	0.92	0.68
40	0.93	0.71
50	0.93	0.7

Table 4. Evaluation results for 500 blog posts, 2000 iterations, $\alpha = 1$, $\beta = 0.1$ and a variable number of topics

7   Conclusions

In this work we have proposed an early version of an automatic evaluation methodology which permits the optimization of topic extraction models for online reputation management using external information not employed by the system. In a preliminar experiment using blog entries and tweets for a bank, we have been able to observe quantitative effects such as a little influence of the α hyperparameter on the final results, the number of iterations which lead to stable results for LDA, or the effect produced by the number of topics.

Our evaluation methodology has still unresolved issues: we do not know yet to which extent the selection of the RDP and NRDP sets bias the results (they are, after all, just a small sample of the full test set, with very precise characteristics). We also need to revise the "temporality" measure to obtain more stable results in our experimental framework.

In any case, the methodology provides a way of testing hypothesis not yet evaluated quantitatively in other studies, such as the effect of including a temporal variable in the model, the possibility of processing time slots independently, the effects of structuring topics hierarchically, etc.

References

1. Hurst, M.: Temporal text mining. In: Proceedings of the AAAI Spring Symposia on Computational Approaches to Analyzing Weblogs. (2006)
2. Subasic, I., Bettina, B.: From bursty patterns to bursty facts: The effectiveness of temporal text mining for news
3. Brooks, C.H., Montanez, N.: Improved annotation of the blogosphere via autotagging and hierarchical clustering. In: WWW '06: Proceedings of the 15th international conference on World Wide Web, New York, NY, USA, ACM Press (2006) 625–632
4. Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A.: Information diffusion through blogspace. In: WWW '04: Proceedings of the 13th international conference on World Wide Web, New York, NY, USA, ACM (2004) 491–501
5. Chi, Y., Tseng, B.L., Tatemura, J.: Eigen-trend: trend analysis in the blogosphere based on singular value decompositions. In: CIKM '06: Proceedings of the 15th

ACM international conference on Information and knowledge management, New York, NY, USA, ACM Press (2006) 68–77

6. Mei, Q., Liu, C., Su, H., Zhai, C.: A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In: WWW '06: Proceedings of the 15th international conference on World Wide Web, New York, NY, USA, ACM Press (2006) 533–542

7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3** (2002) 2003

8. Wang, X., McCallum, A.: Topics over time: a non-markov continuous-time model of topical trends. In: KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, ACM (2006) 424–433

9. Blei, D., Lafferty, J.: Dynamic topic models. In: Proceedings of the 23rd international conference on Machine learning, ACM New York, NY, USA (2006) 113–120

10. AlSumait, L., Barbara, D., Domeniconi, C.: On-line LDA: adaptive topic models for mining text streams with applications to topic detection and tracking. In: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, IEEE Computer Society Washington, DC, USA (2008) 3–12

11. Nallapati, R.M., Ditmore, S., Lafferty, J.D., Ung, K.: Multiscale topic tomography. In: KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, ACM (2007) 520–529