Learning the Importance of Latent Topics to Discover Highly Influential News Items

Ralf Krestel[1] and Bhaskar Mehta[2]

[1] L3S Research Center
Leibniz Universität Hannover, Germany
[2] Google Inc.
Zurich, Switzerland

Abstract. Online news is a major source of information for many people. The overwhelming amount of new articles published every day makes it necessary to filter out unimportant ones and detect ground breaking new articles.
In this paper, we propose the use of Latent Dirichlet Allocation (LDA) to find the hidden factors of important news stories. These factors are then used to train a Support Vector Machine (SVM) to classify new news items as they appear. We compare our results with SVMs based on a bag-of-words approach and other language features. The advantage of a LDA processing is not only a better accuracy in predicting important news, but also a better interpretability of the results. The latent topics show directly the important factors of a news story.

1 Introduction

Online news is a major source of information for many people and has been steadily gaining popularity over printed news. Easy access worldwide and a nearly-realtime availability of new breaking news are big advantages over classical paper-based news distribution. The downside of this success is the huge amount of news items generated everyday. For the common user, this situation presents new challenges, since the volume of news makes it difficult – if not impossible – to keep track of all important events.

The current solution is to look at news aggregator sites like Google News[3]. They offer an automatic clustering of news items into broader categories, and collect news from thousands of sources. This is done using information crawled from online news pages. They also offer a ranking based on the information in the Web. One challenge for such aggregators is to pick the most important stories as they break, and to feature them as soon as they are available. Tradional newspapers employ a team of editors to filter the most important news; automatic approaches need smart algorithms to predict such news. We present in this paper an approach to predict the importance of a news item based solely on the content of an article and its inherent topics. In addition, the main contribution of this work is the generation of an easy to understand representation for important news based on LDA topics. Improvements in accuracy over previous approaches show the effectiveness of our approach.

[3] http://news.google.com/

2 Related Work

Ranking of news is a rather recent discipline among computer science research. But there is a long tradition trying to explain the importance of some events or news. This research is traditionally conducted by communication theorists or journalists. Starting with [1] who introduced *news value* as a measure of importance followed later by [2] who researched which *factors* make a piece of news important. [3] tried to predict importance (*newsworthiness*) of news based on a manual content analysis of experts. With the access to a lot of news online and the overload of a single user, automatic ranking or filtering of news becomes very important.

Most commercial news sites have some mechanism to rank different news articles. Some have experts ranking the news, some rely on social human filtering. Google News provides an automated news aggreation service which also ranks news stories. [4] suggests the use of collaborative filtering and text clustering. Other features include the source, time of publication, and the size of the cluster.

In [5] the goal is to rank a stream of news. This is done by assigning scores to news sources and articles which can also mutually reinforce each other. Another input feature of the algorithm considers the size of the clusters of similar articles. They argue that this size is a good indicator for importance. Considering the streaming characteristic of news, they treat time as a crucial factor. Old news are less important than fresh news in general. The linear complexity of the algorithm makes it applicable for on-line processing and ranking of news items.

Mutual reinforcement of news sites and news stories is also used in [6]. The authors assume that important news are presented on a visually significant spot on the news page. Visual layout is considered one indicator, the second one is the assumption that important news are covered by many news sites. The relation between news sites, news stories and latent events is represented with a graph model. Sites get scores for credibility, events and stories for their importance. These scores are computed via propagation through the graph.

In [7] this approach is modified by changing the graph structure and only considering news and sources and the corresponding relations between them. The use of a semi-supervised learning algorithm is proposed to predict the recommendation strength of a news site for articles on other news sites, which leads to more edges in the graph and yields a better performance for the algorithm. Similarity between articles is measured using a vector space model and the relation between sources and articles are weighted using visual layout information.

All these systems have one common feature; they use information from news pages on the Internet, either taking the number of similar news articles into account or the internal ranking of articles within news pages. The drawback of these approaches is that they give an overview of what news are there and they rank these news items without regarding their intrinsic importance. Since newspapers and news sites have to publish articles even if nothing really important happened, some news stories might get an inflated score, and thus be highlighed. Further, there is an implied dependence on social feedback, or duplication; however, this information is not necessary available when a news item is reported.

3 Importance Prediction of News Articles

Newspapers or online news providers present news in an unpersonalized way. They have editors who pick the most important news stories for their readers. Good newspapers do this in an unbiased and divers way. The newspaper reader will constantly be confronted with new topics, new opinions, and new views, allowing him to broaden his knowledge and to stay informed on important events. The advantages of an edited newspaper are lost if any kind of personalization is employed. Personalization leads to a limited world view that only covers a focused set of topics or events and in the worst case only a certain opinion about a particular topic. A personalized newspaper does not serve the purpose of a general newspaper anymore. There is no more "surprise" for the user. Important topics are filtered out if they do not fit into the previous reading patterns of a user. Instead of getting controverse view points, articles containting the same topics or opinion as the reader has seen before are prefered. Imagine a user in favour of the democratic party who reads a personalized newspaper. She wants to inform herself for the upcoming elections. In the worst case, the personalization system knows about the pro-democratic attitude of the user and only presents pro democratic articles. Diversity or controversal news coverage is not supported.

Nevertheless, filtering of news articles is essential, solely because of the huge amount published every day. But based on the previous paragraphs we argue that it should not be based on personalization but on importance in an unbiased and highly objective way. Even though it is difficult to draw a sharp line between important news and unimportant ones, it seems easy to identify extremely important news and really unimportant news just by looking at the number of news providers covering a given story. Another alternative is to look at user feedback, e.g. click through rates. While these social features are very strong indications, they are often known only after the story has been around for a while. Our aim is to examine news stories without such signals, so that a reasonably accurate prediction can be made as soon as the article is available electronically. The source of a news story can still be used as signal since this information is available at publishing time. However, the news industry today sources news stories from aggregated news agencies (e.g. the associated Press, or Reuters), and clearly not every story can be deemed important. The filtering algorithm we want to devise relies on plain text, and should make the job of human editors easier by picking the most important stories first. This approach can also be imagined in a TV scenario, where transcribed text from the TV audio is run through a classifier and can recognize important stories as they are made available.

From an historical point of view, certain types of events have triggered the creation of many news articles all over the world. Events like the breakout of a war or big natural catastrophies are important in the sense that they get global news coverage. Our goal is to predict the number of newspapers who will pick up a certain topic and thus estimate the importance of this topic.

4 Features for Importance Prediction

Supervised Machine Learning techniques, in particular Support Vector Machines (SVM), need a set of features for each instance they are supposed to classify. For news prediction, extracting certain language features from news articles and using them as input for a SVM is one solution. We will present this approach together with our approach based on LDA feature reduction.

4.1 Importance Prediction from Language Features

Language features have been studied before in the context of news importance prediction [8]. We implemented the same algorithms to compare the performance with our LDA based approach. We analyzed the effectiveness of part-of-speech information and named entities for importance prediction.

Part-of-Speech Information. Part-of-speech information can be very helpful for various classification tasks. In the area of sentiment analysis, e.g., adjectives have been shown to play a superior role over nouns. For topic classification the opposite is true. In this work we want to find out whether a similar result can be claimed for classifying articles into important and non-important ones. We focused on verbs, nouns, and adjectives as labeled by a part-of-speech tagger.

Named-Entity Information. Since we are interested in building a general classifier named entities seem to be counter productive. They describe well a particular instance of a type of important events but if it comes to abstraction they might not help. We experimented with the most common named entities: Persons, locations, organizations, as well as job titles. The hypothesis is that articles mentioning a "President" or the "NATO" might be important.

4.2 Importance Prediction from Latent Features

In the most general form, we represent news with term frequency vectors (TFV). For each news story, we use text from up to 7 different sources, and then combine the document as a TFV. This representation has a very large number of features and the data is very sparse. [8] explored the effectiveness of SVM based importance classifiers on term frequency vectors. While this approach performed well, generalization was difficult due to the sparseness of features and redundancy. In this work, we propose the use of latent factors derived from dimensionality reduction of text as the features for a classifier.

This dimensionality reduction not only generalizes and smoothens the noise but also decomposes the semantics of the text along different latent dimensions. The latent topics are identified using Latent Dirichlet Allocation (LDA) [9]. LDA models documents as probabilistic combinations of topics i.e. $P(z \mid d)$, with each topic described by terms following another probability distribution i.e. $P(w \mid z)$. This can be formalized as

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j), \tag{1}$$

where $P(w_i)$ is the probability of the ith word for a given document and z_i is the latent topic. $P(w_i|z_i = j)$ is the probability of w_i within topic j. $P(z_i = j)$ is the probability of picking a word from topic j in the document. These probability distributions are specified by LDA using Dirichlet distributions. The number of latent topics T has to be defined in advance and allows to adjust the degree of specialization of the latent topics.

An article about the 2008 Presidential elections in the US would then be represented as a mixture of the “election”-topic (say 40%) and an “Obama & McCain”-topic (maybe 30%) as well as some other latent topics (summing up to 30%). Using LDA, we learn a probabilty distribution over a fixed number of latent topics; for the purpose of classification, we treat the probabilities of latent topics as input features.

5 Predicting Importance using SVMs

With a representation for each news story and the knowledge about the *importance* for a certain story , we can use supervised learning techniques to train a classifier. Experiments with different machine learning algorithms have shown that Support Vector Machines (SVM) achieve the best results for this task. The input for SVM are the news stories represented as a mixture of probabilities of latent topics, or the tf-idf weights in the bag-of-words approaches.

Evaluating *importance* seems to be at the first glance a very subjective task. To ensure an objective measure and to clearly differentiate our work from what is known under “personalization” of news, we need a measure that is unbiased. In addition, this measure must provide the posibility of a fully automatical evaluation.

We argue that the number of articles published all over the world for a given news story is a good indicator of its importance. If we choose only a fixed number of sources for measuring importance, a certain bias is introduced nevertheless. Instead of a global importance we might therefore only gain a “western” world view of what is important. By selecting different news sources this bias is eliminated. Further, the number of sources for a story can be easily found from Google news service (we call this *cluster size*).

In our last setup we try to predict this number which is a classical *regression problem*. In the particular case of news importance, the actual number of articles might, however, not be the decisive factor. It is generally enough to differentiate between classes, e.g. unimportant news vs. important ones. Thus we can formalize for our first evaluation setting the task as a *classification task* (labeling the stories as important or unimportant). More fine-grained results are achieved using 4 bins: “extremly important, “highly important”, “moderately important”, and “unimportant”. Therefore the corpus is divided not into two equally sized bins but into bins based on cluster size. This acomodates the fact that there is a majority of “unimportant” news stories in the corpus.

In addition to reporting Accuracy, we focus on Radio Operater Characteristic - Area Under the Curve (ROC-AUC) values [10]. This ensures that we get comparable results even with different sized classes.

For a two-class classification task, where we don’t have an ordering or ranking of the results (e.g. a probability value that an instance belongs to one of the classes) the ROC-AUC value can be computed as: $\textsc{roc-auc} = \frac{1}{2}P_t.P_f+(1-P_f).P_t+\frac{1}{2}(1-P_f)(1-P_t)$ with P_f as the false positive rate ($\frac{\text{false positives}}{\text{false positives+true negative}}$) and P_t the true positive rate.

6 Experimental Results

For our dataset we collected 3202 stories from the Google news service and crawled 4-7 articles per story. These stories were collected over a period of one year (2008).

The input for LDA was generated using GATE [11]; we used LingPipe's [12] implementation of LDA and WEKA [13] for the machine learning experiments. All SVM results were obtained using cross-validation.

To compare our approach to previous approaches for importance prediction [8], we applied the described methods to our corpus. Namely part-of-speech tagging and named entity recognition were used to get an enhanced bag-of-words representation for each news story. With this method, we get ROC-AUC values of up to 0.683 for the two equally sized bins classification compared to 0.728 with our LDA approach. This is an increase of more than 10%.

6.1 Two-Class Classification

In Table 1 (left) the results for filtering the input data making use of part-of-speech information and Named Entity Recognition is shown. The numbers indicate that the preselection of certain word types is decreasing the prediction accuracy. Overall best accuracy is 64.52% achieved using all word types. Nouns tend to have a higher predictive value as e.g. persons. In the following we will compare our results only with the bag-of-words (BOW) approach, since keeping all word types yielded the best results.

Feature	Accuracy	ROC-AUC
all types	64.52%	0.683
verbs, nouns, adj.	63.65%	0.677
nouns	62.90%	0.668
named entities	61.84%	0.645
verbs	58.68%	0.606
adjectives	58.34%	0.608
persons	57.78%	0.598
locations	56.62%	0.589
jobtitles	55.56%	0.581
organizations	55.56%	0.573

No. of LDA Topics	Accuracy	ROC-AUC
50	63.59%	0.682
100	65.58%	0.716
250	64.83%	0.709
500	65.99%	0.720
750	65.15%	0.717
1000	66.27%	0.728
2500	65.87%	0.723

Table 1. Results for using different language features (left) and different number of LDA topics (right) for binary classification on equally sized bins

Using LDA to reduce the number of features improves not only efficiency and interpretability but also accuracy. We evaluated the performance of our algorithm varying the number of LDA topics generated out of the news data. The ROC-AUC values are between 0.682 for 50 LDA topics and 0.728 for 1000 (see Table 1 right). The best accuracy is 66.27%. The higher the number of latent topics, the more specific are the LDA topics.

6.2 Regression Setup

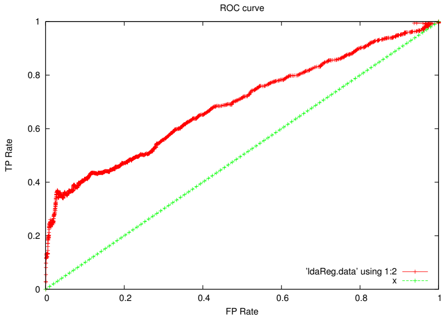

Fig. 1. ROC curve for different thresholds (cluster size) to seperate unimportant and important topics in the regression setup using LDA

The correlation coefficient is 0.47 for using LDA compared to 0.39 when using the bag-of-words approach. Root relative squared error is with 89.14% rather high but still better then using BOW.

Figure 1 show ROC curves for varying the threshold. We therefore do a normal regression and then systematically lower the threshold for a story to be important starting from 1.0. For each threshold we get a false positive rate and a true positive rate. The green line ($f(x) = x$) indicates a random algorithm. Our results are significantly better for both, BOW and LDA.

6.3 LDA Topics

Table 2 shows the three top ranked LDA topics with respect to information gain. A detailed analysis of the model built by the classifier revealed that the first topic (Topic 128) indicates an unimportant news article whereas the other two indicate important news. Since we did this evaluation using 250 LDA topics to represent our documents, some LDA topics contain actually two "topics" (oil, nigeria, indonesia). Other LDA topics indicating importance are e.g.: "McCain, Obama, campaign" or "gas, Ukraine, Russia, Europe".

Topic84			Topic197			Topic128		
Word	**Count**	**Prob**	**Word**	**Count**	**Prob**	**Word**	**Count**	**Prob**
afghanistan	5223	0,120	oil	2256	0,135	able	1803	0,069
afghan	1801	0,041	nigeria	363	0,022	browser	1786	0,068
nato	1732	0,040	company	334	0,020	content	1298	0,049
taliban	1288	0,030	militant	302	0,018	view	1275	0,049
troops	1153	0,027	barrel	280	0,017	style	1206	0,046
country	869	0,020	production	263	0,016	enable	1200	0,046
kabul	709	0,016	crude	246	0,015	sheet	1187	0,045
force	665	0,015	niger delta	240	0,014	css	1136	0,043
security	573	0,013	attack	225	0,013	bbc	1017	0,039
fight	569	0,013	pipeline	211	0,013	internet	720	0,027

Table 2. Top features based on information gain. First two indicating important news; third one indicating unimportance. For each word also the number of occurances in the corpus is displayed, as well as the probability that the word belongs to the topic

7 Conclusions & Future Work

The main goal of this work was to make the importance prediction more accessible in the sense of easy interpretable results. We have shown that LDA can achieve this. It is possible to identify general news events, e.g. the war in Afghanistan, and predict the importance of future articles dealing with this topic. Also general events like elections were identified by LDA and help to predict the coverage of other future elections in the media. In conclusion, we explored a new approach to the problem of finding important news as a classification problem using LDA topic mixtures. The results show that accuracy is better than state-of-the-art bag-of-words results. We can find important articles with an accuracy considerably better than random, and around 10% better than previous approaches.

The biggest open issues concerns time. Our current approach has no time dimension. Incorporating the temporal aspects of news articles is decisive to improve accuracy further. Not only because "nothing is older than yesterday's news", but also because of the interdependence of news stories. A news story might not be important only based on its content but also because there was another news story related to it. For future work we try to involve these temporal aspects of news. Extending the LDA implementation to consider a time dimension might be neccessary to achieve this. We also try to incorporate more knowledge from the domain of journalism where research on importance factors of news articles has been carried out. A detailed analysis of mentioned numbers or dates with articles might also improve accuracy.

References

1. Lippmann, W.: Public Opinion. Harcourt, Brace and Company New York (1922)
2. Østgaard, E.: Factors influencing the flow of news. Journal of Peace Research **2** (1965)
3. Kepplinger, H.M., Ehmig, S.C.: Predicting news decisions. an empirical test of the two-component theory of news selection. Communications **31**(1) (April 2006)
4. Das, A.S., Datar, M., Garg, A., Rajaram, S.: Google News Personalization: Scalable Online Collaborative Filtering. In: Proc. of the 16^{th} World Wide Web Conference, ACM (2007)
5. Corso, G.D., Gullí, A., Romani, F.: Ranking a stream of news. In: Proc. of the 14^{th} international conference on World Wide Web, ACM (2005)
6. Yao, J., Wang, J., Li, Z., Li, M., Ma, W.Y.: Ranking web news via homepage visual layout and cross-site voting. In: Advances in Information Retrieval, ECIR'06, Springer (2006)
7. Hu, Y., Li, M., Li, Z., Ma, W.Y.: Discovering authoritative news sources and top news stories. In: AIRS'06. Volume 4182 of Lecture Notes in Computer Science., Springer (2006)
8. Krestel, R., Mehta, B.: Predicting news story importance using language features. In: Proc. of 2008 IEEE / WIC /ACM International Conference on Web Intelligence, IEEE (2008)
9. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3** (January 2003) 993–1022
10. Provost, F., Fawcett, T., Kohavi, R.: The case against accuracy estimation for comparing induction algorithms. In: Proc. of ICML'98, Morgan Kaufmann (1998)
11. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A framework and graphical development environment for robust NLP tools and applications. In: ACL'02. (2002)
12. Alias-i: Lingpipe 3.7.0. http://alias-i.com/lingpipe (2008)
13. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. $2^{n}d$ edn. Morgan Kaufmann (2005)