Searching and Finding Strikes in the New York Times

Iris Hendrickx, Marten Düring, Kalliopi Zervanou and Antal van den Bosch

Centre for Language Studies, Radboud University Nijmegen, The Netherlands
E-mail: {I.Hendrickx, M.During, K.Zervanou, {A.vandenBosch}@let.ru.nl

Abstract

The huge digitization step that archives, publishers, and libraries are currently undertaking enables access to a vast amount of information for historians. Yet, this does not necessarily make life easier for historians, as the main problem remains how to find relevant sources in this sea of information. We present a case study demonstrating how automatic text analysis can aid historians in finding relevant primary sources. We focus on strike events in the 1980s in the USA. In earlier work on strikes, researchers did not have at their disposal a full and comprehensive list of major strike events. Existing databases of this kind (e.g [19, 22]) are the result of intensive manual work and took years to build. Natural language processing (NLP) tools allow for faster assembly of datasets of this kind on the basis of collections of free texts that contain the information that should be in the database. We aim to construct a database of events using a digital newspaper archive and unsupervised NLP methods such as Latent Dirichlet Allocation (LDA) and clustering techniques to group together newspaper articles that describe the same strike. We study the effect of different feature representations, such as simple bag-of-words features, named entities, and time stamp information. We evaluate our results on a manually labeled sample of news articles describing a small set of strikes.

1 Introduction

The time period of the 1980s is a interesting period to study social unrest as this period saw great social, economic, and general change. The early 1980s were marked by a global recession in the industrialized countries. Many multinational corporations migrated their factories to emerging countries in Asia and Mexico leaving behind many unemployed workers. In the United States, the Reagan administration strived for free market economy and tax cuts to stimulate economic growth [15].

In this study we focus on strike events as proxy indicators of social unrest. We view a strike event as a labor action by a (significantly large) group of workers at a certain company, governmental body or specific sector that halted their work for a clear motivation, such as better working conditions, or a higher salary.

We use automatic unsupervised techniques and an online available newspaper archive as our source to find out which individual strikes took place in the 1980's. In particular, our objective is to cluster news articles based on the strike they refer to in order to seed a database of strike events. In our case study, we work with the online archive of the New York Times (NYT)[1], one of the largest and most influential daily newspapers in USA. The NYT offers an online search interface[2] to their archive that covers newspapers from 1981 to now. To facilitate the search, the articles have been manually labeled by NYT with metadata and controlled keywords (facets) such as locations, topics, and normalized names. One of these keywords is 'STRIKES', which trivially allows to retrieve all strike-related news articles from the 1980s (assuming the keyword is assigned correctly and completely). In total there are 5,987 articles on strikes in this period which form the basic set in our study. Still, the facet does not give information on which articles talk about the same strike. To identify this is a challenging task, as all strike articles are covering the same topic and therefore are similar to each other. We can expect many of these articles to cover sub-events such as picketing, negotiations between worker unions and company directors, or demonstrations. To discover how many different individual strike events took place, we need to focus on who was involved in these strikes, where they took place, and in which time period.

Our approach operates at the document level. Our task can be seen as a type of text clustering where the goal is to detect a latent group structure [18]. Note that our approach is different from a subcategorisation frame-filling approach [1] or a template-filling approach [13] to event detection, where the aim is to find an event trigger, usually a verb, and some slot or template fillers, such as agent, goal, and location at the sentence level. In our text clustering definition of an event at document level, the particular variables (who, what, when) are not explicit in the event representation. However, these variables are crucial in distinguishing one strike from another and are therefore important features in the document representation. Even though we cannot know beforehand which company names or sectors to expect, we do know that names of organizations, such as companies and labor unions will play an important role in establishing that two articles are describing the same event. Time also plays an important role, as we can expect to find news articles on the same strike when they are close in time, and when they mention the same start date of the strike, its duration and possibly an end date.

In this investigation, we use off-the-shelf NLP tools that operate on relatively simple document representations to automatically cluster documents describing the same strike. We compare two unsupervised techniques to achieve this goal. First,

[1] `http://www.nyt.com`

[2] NYT API, version 1: `http://api.nytimes.com/svc/search/v1/article`

we use *sequential Information Bottleneck* (sIB) [20], a clustering method that efficiently casts bottom-up agglomerative clustering as a sequential clustering process. This clustering algorithm has been shown to be successful in document clustering. Second, LDA [2] has become a popular method for event clustering (e.g. [24, 5]). In LDA, a document is seen as a probability distribution over a set of topics, where a topic is a probabilistic distribution over a set of words. We assume that documents about the same strike event share a similar probability distribution and the same topic will be assigned to the articles about the same event.

Note that applying a supervised technique is not feasible for this problem. Manually labeling a subset of the strike articles is not likely to be representative for detecting other individual strikes as the documents are so much alike and the unique features that will distinguish one strike from another are not expected to re-occur.

In the rest of this paper, we first present related work in Section 2. We then detail the experimental setup of our investigation in Section 3 and discuss our results (Section 4). Finally, we conclude in Section 5 with the principal findings of this study and our plans for future work.

2 Related work

A study similar to ours was carried out by Yang et al.[24]. They also aim to support historical research by applying topic modeling on a collection of historical newspaper articles. They however take an explorative and open approach and they study to what extent automatically generated topics (represented as a set of keywords) reflect historical trends. They use an expert historian to interpret the found topics. In our approach, we rather focus on a specific type of event and use the topics as document labels to find documents describing the same strike event.

De Smet and Moens [5] studied the representation of news articles for event clustering. They focus on short term events from Wikinews, and their data set had only a few relevant articles per event. They show in their work that a complex methods such as LSA or LDA did not outperform the classic tf*idf-weighted [17] bag-of-words approach on the task of event detection.

In the work of [9] LDA is applied to clustering texts into general categories such as books, business, fashion, etc. They focus on dealing with short messages as they work with tweets. [23] also apply their own proposed algorithm based on Wavelet-based signals and LDA to event detection in tweets. In their manual evaluation of the results, they conclude that LDA topic models are difficult to interpret and evaluate.

3 Experimental setup

In our experimental setup, we follow the approach taken by [5] and compare a tf*idf-weighted bag-of-words approach to LDA topic modelling. We first perform

a baseline experiment with a feature representation based on content words only. In a second experiment, we use additional features to explicate and emphasize aspects that are likely to distinguish one strike event from another: persons, locations, organizations and time stamps.

Our data set consists of the 5,987 articles that were labeled as relevant to 'STRIKE' in the NYT online archive in the the period 1980–1989. Each article is automatically tokenized, part-of-speech tagged and labeled with named entity (NE) information and time information using the Stanford CoreNLP tools [21, 7].

The Stanford POS-tagger achieved an accuracy of 97.2% on the well-known WSJ test set [21]. The Stanford NE tool obtains an F-score of 87% on the CoNLL 2003 shared task material [7]. Both test sets consist of newspaper text similar to the data that we are working with in our experiment.

We retain only content words (i.e. nouns, adjectives, verbs, adverbs) as the basis for feature representation. For the clustering, we use a bag-of-words feature vector representation. Each word receives an importance weight by computing its tf*idf score [17]. We represent each article as a weighted word vector using the 20,000 most frequently occurring content words in the data set.

For topic modeling the sequential order of the words in the article is kept, as topic modeling is based on word co-occurences. We assume that articles describing the same strike event share a similar probability distribution and the same topic will be assigned to the articles about the same event. Therefore, we assign the topic label with the highest score to each article, using the topics as class labels. We do not implement a frequency cut-off for the LDA feature representation. We apply the LDA algorithms as implemented in the Mallet toolkit [11]. For LDA, we use Gibbs sampling and asymmetric Dirichlet priors on document-topic distributions, as it was found to lead to better results [12]. In their study, MacCallum et al. [12] also showed that it is generally better to set the number of topics too large than too small. With a good topic model, the superfluous topics will have few entries, and the overall distribution of topics will still be good. In our experiments, we tested a range of 100, 150, 200, and 300 for the number of topics.

For the siB clustering algorithm, we use the Weka toolkit implementation [8] and 25 optimization iterations. As we do not know how many individual strike events actually occur in our full data set, we tested a same range of numbers of clusters as for LDA. As both unsupervised methods take random initializations, we repeated each experiment ten times with different random seeds and report the average over these ten runs.

In a second round of experiments, we add the named entities and time expressions (dates and duration) as predicted by the Stanford NLP tools as extra features to the feature vector representation. For example, the named entity string *Transport_Workers_Union* was added to the vector as a new feature, in addition to the separate words *Transport*, *Workers*, and *Union* that were already present in the feature vector. Numbers were excluded from the bag-of-words vector in the first experiment, but are now kept when part of a time expression such as *1984* or *January_1980*. Replacing individual tokens by longer strings instead of adding them

would lead to more sparsity. This way we hope to profit from the more complex and meaningful named entities in combination with the robustness of simple word frequencies.

3.1 Evaluation

For the evaluation of our approach, we manually labeled a sample of NYT articles with strike event information. Many recent studies on event detection either used a fully labeled data set (for example [5, 9, 4]), or they manually evaluated their results afterwards (for example [24, 23]). Here we chose to label a small sample beforehand in the following way. First, one of the authors, a historian, created a small list of eight strike events that occurred in the 1980s using different sources (for example [6, 19, 14, 3]). Next, he verified that all articles in the NYT archive describing these eight events were found. In total, 299 articles were linked to the eight strikes. The manual sample exhibits unevenly balanced amounts of articles per strike. The '1981 Major League Baseball' strike alone is linked to 100 relevant articles, while for the 'Chicago Tribune newspaper' strike in 1986 we only found 2 articles in the NYT. This also exemplifies a weak point in our current study: we aim to create a list of strikes in the USA in the 1980s only based on one source, the NYT. We do assume that the NYT will report on all major strikes in the country, but it is likely that a strike in Chicago is documented more extensively by regional newspapers. We run the unsupervised techniques on the full set of almost six thousand articles, and we compute recall, precision and F-scores on the manually labeled subset of 299 articles.

Both unsupervised techniques that we have applied (i.e. LDA topic modeling and siB clustering), assign an arbitrary label to each cluster, and we need to match these labels to the true manual individual strike event clusters. For every true event cluster, we check which arbitrary label was predicted most times, and we mark this label as corresponding to the true label. Once a true label is chosen as matching to a given cluster, it cannot be re-assigned to any other cluster. We computed recall, precision and F-scores on the found correspondences.

4 Results

In this section, we present the results of our experiments starting with the baseline experiments with single word features as input for LDA and the tf*idf weighted word vectors for siB clustering. We present micro-averages over the 299 documents that cover 8 strike events, averaged over 10 random initializations. The results of both algorithms with the varied number of topics/clusters can be seen in table 1 and is depicted in the left-side figures 1a and 2a. The siB algorithm performs better than LDA. In alternating the desired number of topics/clusters, we observe a clear trade-off between recall and precision. As illustrated in the figure for LDA (2a), there is an increase in precision and a decrease in recall, when we

	sIB			LDA		
# topics	recall	prec	Fscore	recall	prec	Fscore
100	70.2	51.8	59.6	64.6	40.6	49.9
150	59.7	56.2	57.9	63.8	45.6	53.1
200	49.7	61.9	55.3	61.6	48.7	54.3
300	38.7	70.3	49.9	59.4	52.1	55.5

Table 1: Baseline experiments. Micro averages of siB clustering and LDA, averaged over 10 randomly initializations. Scores were computed on the labeled sub set of 299 articles belonging to 8 different strike events.

increase the number of topics. A similar trend with steeper curves is also indicated for the siB clustering experiments (figure 1a). It is interesting to observe that while for both algorithms show this same trade-off, the overall F-score of siB decreases when increasing the number of clusters, while for LDA the F-score increases.

In table 2 and 3 we zoom in on the results for the individual strike events for those cases that have shown the highest F-score in table 1 (100 clusters for sIB and 300 topics for LDA).

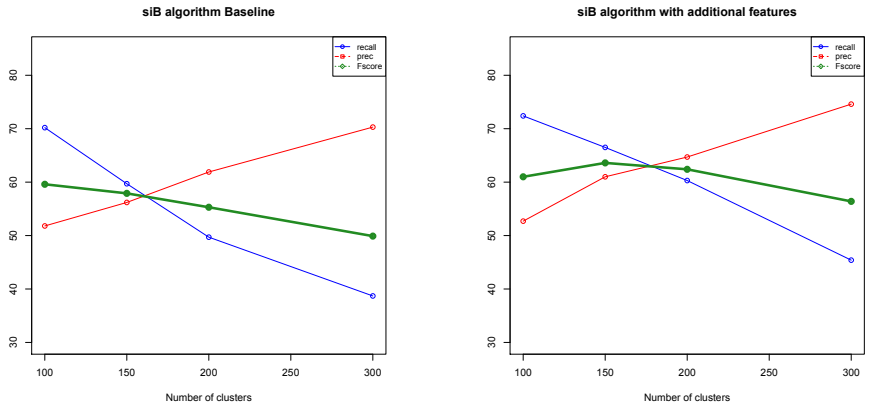
# Art	Strike event	Recall	Prec	F-score
100	Baseball League	46.1	38.3	41.8
68	Austin Hormel	92.5	88.4	90.4
52	Pittston Mine	86.4	68.4	76.2
41	Writers 1981	88.3	41.6	56.5
17	Writers 1988	5.9	1.9	2.8
14	T. W.A.	89.29	21.1	34.0
5	Arizona Copper	100.0	10.1	18.3
2	Chicago Tribune	60.0	2.5	4.8

Table 2: Results for each of the individual strike events with clustering algorithm siB with 100 clusters and 25 iterations (averaged over 10 random initializations).

In Table 2 we list the recall, precision, and F-scores for each of the eight strike events for the results of the clustering algorithm siB with 100 clusters and 25 iterations (averaged over ten random initializations).The rows in the table are ordered on event cluster size. We observe large differences between the scores on the different strike events. On the largest cluster about the Baseball League strike an F-score of only 41.8% is achieved, while the Austin Hormel meat packer strike achieved the best F-score of 90.4%. We see that low scores are obtained for the strikes represented by only a few relevant documents. We observe a very low F-score of 2.8% for the Writers strike in 1988. The explanation of this is that the algorithm did not

# Art	Strike event	Recall	Prec	F-score
100	Baseball League	26.7	39.9	31.4
68	Austin Hormel	90.3	77.3	83.3
52	Pittston Mine	72.1	67.5	66.5
41	Writers 1981	79.0	47.3	57.1
17	Writers 1988	17.1	14.1	13.6
14	T.W.A	76.4	36.3	47.7
5	Arizona Copper	98.0	26.5	40.7
2	Chicago Tribune	50.0	4.1	7.3

Table 3: Results for each of the individual strike events with LDA with 300 topics and 10 optimization iterations (averaged over 10 random initializations).



(a) (b)

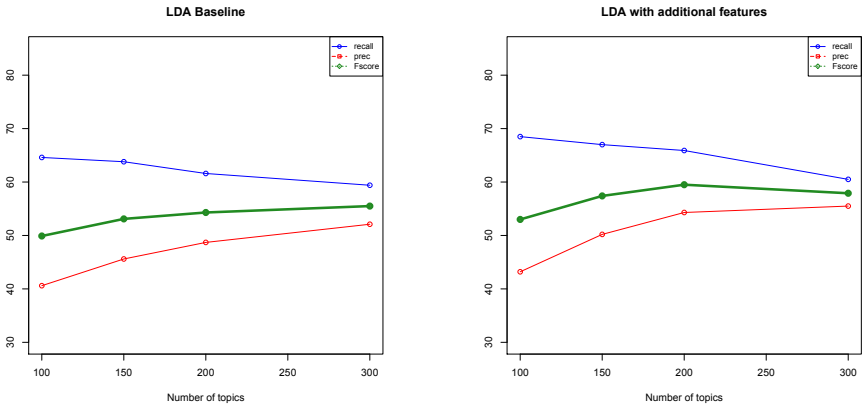
Figure 1: Micro averages of siB clustering. The baseline experiments are shown in (a) and experiments with the additional named entities and time information in (b).

succeed to distinguish the two strikes about writers: they were clustered together and were assigned the same cluster label.

In Table 3 we list the results for the same strike events obtained with LDA topic modeling with 300 topics. In general, we observe lower F-scores than with sIB for the larger event clusters and higher F-scores for the strike events covered by only a few news articles due to an increase in precision.

In the second round of experiments, we exploit additional features for named entities and timestamps in the articles. These features indeed help, as we observe an increase in the overall performance of both LDA and siB clustering as shown in figures 1 and 2.

For siB (figure 1) a clear improvement in precision and a slight decrease in recall can be observed when adding the new features. In figure (1b) sIB attains a



(a) (b)

Figure 2: Micro averages of LDA. The baseline experiments are shown in (a) and experiments with the additional named entities and time information in (b).

maximal F-score of 63.6% with 150 clusters and combines a precision of 61% and a recall 63.6%.

For LDA in figure 2, both recall and precision improve with this new feature set. The run with 200 clusters in figure (2b) has the highest F-score of 59.5%, a recall of 65.9% and 54.3% precision.

4.1 Qualitative analysis

In this section, we look at some examples to see which words are chosen by the LDA method to represent the individual clusters. This gives us some insight to what extent the techniques were indeed taking the important words as their most important features.

In Table 4 we show an example of the top words for each topic produced by LDA for three individual strikes [3]. As shown in table 3, the average results for the meatpackers strike at Hormel were rather accurate: an F-score of 82% with a recall of 91%. When we look at the top words of the most assigned topic label, we can see that all words in this topic are indeed relevant and specific. The strike took place in Austin, Minnesota and Rogers and Guyettte were union members and spokesmen in this strike against Geo A. Hormel & Company in 1986.

Most of the top words for the topic assigned to the articles about the strike event at Trans World Airlines in 1986 are indeed relevant, but this topic does not exclusively describe the T.W.A. strike, because the terms referring to Pan Am and Continental point to other companies not involved in this particular strike. This mix-up of several strikes related to airline companies explains the low precision (9.8%) and higher recall (64.3%) for this particular strike cluster.

[3] we show one example from the 10 random initializations with 200 topics.

Strike of meatpackers at Geo A. Hormel & Company in 1986
hormel plant company austin workers local p strikers union rogers meatpackers food minn parent commercial united plants geo guyette a
Strike at Trans World Airlines in 1986 of flight attendants and machinists
pilots airline flight attendants united airlines pan am company association flights continental american line international hired machinists mechanics carrier percent
The 1981 Major League Baseball
owners players miller grebey baseball kuhn committee moffett relations league commissioner negotiations bargaining owner player steinbrenner negotiating donovan ray labor
players owners free compensation player agent team agents club clubs pool baseball association proposal miller league teams grebey signing agency
baseball players league mets game yankees season stadium play team year yankee home major ball manager club pitcher sox games

Table 4: Examples of the topic models for three manually labeled strikes

For the Major League Baseball event we see the opposite happening: this strike is not covered by one main topic, but by three topics (topics 149,37,194), each covering about 25% of the articles. These topics are clearly all relevant and closely related to each other, because the terms *owners players baseball league* occur in all three topics.

LDA is clearly capable of finding topics that model the individual strikes. The top words in the topics point to the relevant sectors, worker groups (e.g. meatpackers, players, mechanics), companies, organizations, spokespersons and actions (e.g. negotiating, signing) that were parts of these strikes. However, an absolute one-to-one correspondence with the individual clusters was not found for all of the clusters in the small manually labeled sample.

5 Conclusions

In these experiments, we addressed the challenging task of clustering strike-related news articles into unique strike event clusters using unsupervised techniques and shallow word and named-entity features. Our experiments have shown that both LDA and the siB algorithm are capable of detecting those specific and relevant word features that distinguish individual strike events. For events with a smaller media coverage, we obtained lower results due to a loss in precision as multiple events are assigned to the same cluster.

The results in this case study indicate that these unsupervised techniques cannot detect individual strike events fully automatically. The results are not good enough for that. Still, it can be expected that the automatic clustering techniques will

reduce the work of the historian. Large events can to a large extent be detected automatically, and many articles can be automatically assigned correctly to single events, so that manual work can concentrate more on the long tail of incorrectly clustered articles.

As a next step we would also like to repeat this study for older historical documents. We refer to the work of [16]. They compare four named-entity extraction systems including the Stanford NE tool [7] applied to historical text material collected by applying optical character recognition to images of typed holocaust testimonies. The Standford NE tool performs best but with low scores between 44% and 60% F-score. As our method relies heavily on the performance of the named entity detector, this gives us an indication of what we can expect if we change to older newspapers. In such case retraining the NE tool will be necessary.

In our current study we treated LDA's topics as clusters. In future work, we would like to experiment with the topics that LDA produced as features instead of cluster labels, similar to how Lin and Hovy use topic signatures for summarisation [10]. Since LDA assigns multiple topic labels to each document, this rich topical representation could be the input of a subsequent unsupervised clustering approach, e.g. with sIB. This way we may combine the different strengths of the two methods.

6 Acknowledgements

We are grateful to the New York Times for sharing their articles for via the Times Developer Network. This work has been carried out within the framework of the Digging into Data project ISHER[4].

References

[1] Roberto Basili, Cristina Giannone, and Diego De Cao. Learning domain-specific framenets from texts. In *Proceedings of the ECAI Workshop on Ontology Learning and Population.*, Patras, Greece, 2008.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[3] Aaron Brenner, Benjamin Day, and Immanuel Ness. *Encyclopedia of Strikes in America*. ME Sharpe, 2011.

[4] Yan Chen, Hadi Amiri, Zhoujun Li, and Tat-Seng Chua. Emerging topic detection for organizations from microblogs. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 43–52. ACM, 2013.

[4]Integrated Social History Environment for Research – Digging into Social Unrest: `http://www.diggingintodata.org/Home/AwardRecipientsRound22011/ISHER/tabid/196/Default.aspx`

[5] Wim De Smet and Marie-Francine Moens. Representations for multi-document event clustering. *Data Mining and Knowledge Discovery*, 26(3):533–558, 2013.

[6] Steve Early. Strike lessons from the last twenty-five years: Walking out and winning. In *Against The Current*, volume 124. 2006. http://www.solidarity-us.org/current/node/113.

[7] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics, 2005.

[8] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11(1), 2009.

[9] Liangjie Hong and Brian D Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88. ACM, 2010.

[10] Chin-Yew Lin and Eduard Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 495–501. Association for Computational Linguistics, 2000.

[11] Andrew McCallum. MALLET: A machine learning for language toolkit. 2002. http://mallet.cs.umass.edu.

[12] Andrew Mccallum, David M. Mimno, and Hanna M. Wallach. Rethinking LDA: Why Priors Matter. In *Advances in Neural Information Processing Systems*, pages 1973–1981, 2009.

[13] MUC-7. Muc-7 coreference task definition. version 3.0. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1998.

[14] Bradley Nash Jr. *The Crises of Unions in the 1980s*. PhD thesis, Virginia Polytechnic Institute and State University, 2000.

[15] William A. Niskanen. Reaganomics. concise encyclopedia of economics. *Library of Economics and Liberty*, 1992.

[16] Kepa Joseba Rodriquez, Mike Bryant, Tobias Blanke, and Magdalena Luszczynska. Comparison of named entity recognition tools for raw OCR text. In *Proceedings of KONVENS 2012*, pages 410–414. ÖGAI, 2012.

[17] Gerard Salton and Michael McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, 1983.

[18] Fabrizio Sebastiani. *The Encyclopedia of Database Technologies and Applications*, chapter Text categorization. Idea Group Publishing, 2005.

[19] Beverly J. Silver. *Forces of labor: workers' movements and globalization since 1870*. Cambridge University Press, 2003.

[20] Noam Slonim, Nir Friedman, and Naftali Tishby. Unsupervised document classification using sequential information maximization. In *Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 129–136, 2002.

[21] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics, 2003.

[22] Sjaak Van der Velden. *Stakingen in Nederland, Arbeidersstrijd 1830-1995*. Stichting beheer IISG/NIWI, 2009. Available online at: http://www.onvoltooidverleden.nl/fileadmin/redactie/Velden/Stakingen_in_Nederland.pdf.

[23] Jianshu Weng and Bu-Sung Lee. Event detection in twitter. In *Proceedings of the fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[24] Tze-I Yang, Andrew J. Torget, and Rada Mihalcea. Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–104. Association for Computational Linguistics, 2011.