Strikes that never happened

Text mining in historical data

Martha van den Hoven

ILK Technical Report Series 10-05

August 2010

ILK Research Group
Tilburg Center for Cognition and Communication
School of Humanities
Tilburg University
P.O. Box 90153
NL-5000 LE Tilburg, The Netherlands

Abstract

This master thesis is about information retrieval in historical archives. It reports on research that has been carried out as part of the HiTiME project, a collaboration between the International Institute of Social History (IISH) in Amsterdam and Tilburg University. The strikes database of the IISH forms the entry point. The question is if a method can be set up to find related newspaper articles, in the digital newspaper archive of the Royal Library. A query model has been made to automatically find articles relating to a particular strike, and the results have been ranked.
Using the information and articles that can be found about actual strikes, an analysis is made of the articles written about the unrest and conflicts that go on before the real strike breaks out. These findings are then used to try and search for conflicts that were about to burst into strike, but for some reason never did: the strikes that never happened.

Contents

1. Introduction 6

2. Tools and methods 9

2.1. TiMBL 9

2.2. Tadpole 9

2.3. Boolean retrieval 10

2.4. Similarity scoring 11

2.5. Mean average precision 12

3. Data 13

3.1. Strikes 13

3.2. Newspapers 15

4. Searching for newspaper articles 18

4.1. Querying for articles 18

4.2. Ranking the results 21

5. Finding strikes that never happened 24

5.1. Strike prelude patterns 24

5.2. Finding actual strike threats 26

5.3. Ranking the results 29

6. Conclusions and recommendations 31

7. Bibliography 34

Appendix A Feature words 36

Appendix B Prelude candidate frequencies per week 38

List of tables

Table 1 Important attributes of a strike 14
Table 3 Quantity and precision of retrieved newspaper articles 20
Table 4 Results cosine similarity ranking strike articles 22
Table 5 Results cosine similarity ranking with augmented record 23
Table 6 Strikes used to find preliminary articles 25
Table 7 Top 10 feature words in prelude vs. non-prelude classification 26
Table 8 Results of categorisation possible prelude articles 28
Table 9 Results cosine similarity ranking prelude articles 30
Table 10 Gain ratio of feature words in prelude vs. non-prelude classification 36

List of figures

Figure 1 Example of a strike record 14
Figure 2 Percentage of candidate prelude articles per week 1912 38
Figure 3 Percentage of candidate prelude articles per week 1926 38
Figure 4 Percentage of candidate prelude articles per week 1938 38

1. Introduction

In November 2009 the HiTiME project started. This project is a cooperation between the International Institute of Social History (IISH) in Amsterdam and the Tilburg center for Cognition and Communication (TiCC). The goal of the project is the creation of a "historical web of domain knowledge" (HiTiME, 2010), using several of IISH's archives and databases that are digitally available. The end product should be "an associative network of concepts (historically relevant entities) and their relations, linked to a timeline". In the HiTiME project technologies such as text mining and automated knowledge base building will be used to develop a "generic reusable toolkit" that can be used for historical research purposes.
A project such as HiTiME is a great opportunity for the participants. The language technologists have tools and techniques that they would like to try on substantial amounts of real life data, and the historians have these heaps of data that they would like to make accessible, and make sense of. The research of this master thesis consists of a small exploration of what can be done when the tools and the data are brought together.

Several resources are made available to the HiTiME project by IISH, some of which are: a database of biographies of important figures in the Dutch socialist movement; a collection of letters written by Bakunin, the Russian anarchist; the HISVAK database, containing information on all Dutch trade unions that ever existed; a database with information regarding organisations by and for post-colonial migrants; the HISCO database of historical occupations; and a strike database, containing strikes that occurred in the Netherlands.
With so much and such diverse data, there is a wealth of possibilities for research. For this thesis I have chosen to use the strikes database as a starting point for my explorations. This database has been set up and filled by Sjaak van der Velden in the context of his PhD research (Van der Velden, 2000). There are over 16,000 records in the database describing strikes, lockouts and other actions, with the earliest record dating from 1372, regarding a strike of fillers in Leiden. The majority of the strikes recorded in the database occurred between the years 1900 and 1940.

It could be useful for historians if, while browsing the strikes database, they could find more information related to a particular strike. Now the Dutch Royal Library has been digitizing many newspapers over the years, in the project Databank Digitale Dagbladen[1] (Databank of Daily Digital Newspapers). When this website was officially launched on May 27th 2010, one million newspaper pages were put online, to be browsed and searched. The goal is to have eight million pages online by the end of the project. In this research I will try to find a way to automatically retrieve newspaper articles that report about a strike, using information from the strike record.

That this type of research can be of use to historians is demonstrated by Beverly Silver in 'Forces of Labor' (Silver, 2003). This work describes the dynamic of labour movements, starting from the year 1870. In order to describe these waves of labour unrest, large data collections are needed. As Silver states, "Long-term time series of strike activity - the most commonly used indicator of labour unrest - exist only for a handful of countries. [...] Data collections covering non-strike forms of labour unrest are even more rare, yet they are important to the overall construction of a map of labour unrest." In order to overcome this gap in available data, Silver decided to create a new collection, the so-called World Labour Group database. This database is constructed by manually going through the indices of The Times (London) and the New York Times, and selecting relevant mentions. The work that has gone into this undertaking is immense. If there is a way to use automatic information retrieval techniques to do the same work, this type of historical research could benefit hugely.

The next step that we will try to take in this research is to zoom in on the newspaper articles in the period leading up to a strike. There can be mention of talks between employers and employees, possible conflicts, unrest, ultimatums. Taking a certain strike as the point of entry to the newspapers, we can try to find articles mentioning this possible preamble to the strike. But there must have been similar processes where the outcome of the conflict was not a strike but agreement, or maybe a strike has been avoided in some other way. It could be of significance to historians to find these "strikes that never happened". If it is interesting to know why a conflict broke into strike, it would be just as interesting to know why another conflict did not.

---

[1] http://kranten.kb.nl/

In order to find these otherwise resolved conflicts they are assumed to be *counterfactual* strikes, in that the prelude to the non-strike is assumed to be of the same structure as the prelude to the strike. For the people that were involved in the conflict when it happened, the eventual outcome of it (strike or no strike) was not a given as it is now looking back. As Ferguson (1997, p. 88) states: "[...] what actually happened was often *not* the outcome which the majority of informed contemporaries saw as the most likely: the counterfactual scenario was in that sense more 'real' to decision-makers at the critical moment than the actual subsequent events."

Of course, the interpretation of the counterfactual strikes that might be discovered is for the historians. For me, the path to finding them is the subject of my research.

2. Tools and methods

In this chapter some of the language technology tools from the toolkit mentioned earlier will be described. All tools and technologies mentioned in this chapter are used in the research.

2.1. TiMBL

The first tool to describe is memory-based learner TiMBL. This is a system that is designed and built at Tilburg University and the University of Antwerp (Daelemans et al., 2009). TiMBL performs classification tasks using the *k*-Nearest Neighbour algorithm, where every new unclassified instance is assigned the class of the majority of its *k* nearest neighbours, that is the *k* instances that are most similar in features to the new unknown instance. So instead of creating rules to predict the outcome in certain new situations, TiMBL uses analogy to previous known situations.
TiMBL evaluates the features of every instance with regards to the amount of information they reveal about the class of the instance. The features can be scored using several scoring mechanisms: Information Gain (IG), Gain Ratio (GR), chi-squared statistic and shared variance. In this research the former two are used. The IG weight looks at each separate feature and estimates how much information that feature contributes to determining the correct class of the instance. It is the difference between the uncertainty about the class in a situation without knowledge of that feature's value, and a situation with that knowledge. A problem with IG is that it tends to overrate features with many values. Using Gain Ratio solves that problem, by dividing the IG value by the split info, the entropy of the feature values (Daelemans & Van den Bosch, 2005, p. 30).

2.2. Tadpole

Tadpole is a morphosyntactic tagger and dependency parser, that, like TiMBL, is developed at the University of Antwerp and Tilburg University (Van den Bosch et al., 2007). It uses TiMBL for the classification tasks it executes to do the tagging and parsing tasks. When fed with Dutch text, Tadpole tokenises it, splitting the sentences into linguistic units. A token is usually one word, but can also be more, for example

'Tilburgsche Courant'. These tokens are then, depending on what the user needs, given part-of-speech tags, and morphologically analysed. Part-of-speech tags are labels like Noun, Adjective, and Verb. The morphological analysis means the word is split up in the different meaningful parts: *gesteld* becomes [*ge*][*stel*][*d*], with *stellen* as lemma, the canonical form of the word. Finally the dependency parser creates a hierarchical tree structure of each sentence.

The component of Tadpole that is most useful to this research is the part-of-speech tagger. This module gives a specific tag to proper names, such as people, locations and organisations. These so-called named entities are pivotal to this research: it is certain people who organised strikes in certain places, and it is essential for the retrieval of information that they are correctly identified.

2.3. Boolean retrieval

We also need tools for retrieving information. Boolean retrieval is a classic method of information retrieval. Using Boolean queries documents can be found on the basis of whether the search terms, or combinations thereof do exist in the document or not. There is no scoring or ranking of the results other than "yes, this document is a correct result for the query" or "no, this document does not apply". Now with the Boolean operators, AND, OR and NOT, a great variation of queries can be manufactured. But in these queries there is always a trade-off between precision and recall, where precision is the percentage of correctly retrieved results in the total result set, and recall is the percentage of retrieved results out of the total of results that should have been retrieved.

Lee & Fox (1988) argue that the use of AND in queries results in a high precision, at the cost of recall, where the use of OR does the opposite, yielding high recall at the cost of precision. A query searching for 'A and B' will deem all documents containing only A or only B just as irrelevant as documents containing neither. And a query searching for 'A or B' considers documents containing only one of the search terms just as relevant as documents containing both. They make a case for extending the Boolean system with possibilities of assigning weights to the search terms.

Salton et al. (1983) made a similar plea for loosening the interpretation of OR and AND in Boolean systems, as a bridge between the strict Boolean logic and the total absence of structure in vector processing systems.

Manning et al. (2008, pp. 308) do point out that hand-written rules in the form of Boolean queries can achieve a high accuracy (a combination of precision and recall) when they are developed by humans with knowledge of the domain, tuning their queries. They also add that more sophisticated query languages would achieve even better results.

Such an extended query language could involve things like fuzzy search, where the terms in the retrieved documents are allowed to differ one or more characters from the search term. Wildcards could be introduced. Stemming of terms, so that a search term *strike* would also match *strikers* and *striking*. Another possible extension of the Boolean system could be a function like nearness, where search terms have to be found within a certain distance of each other.

2.4. Similarity scoring

When results, or in the case of this research, newspaper articles are retrieved by some sort of querying as described above, it is useful to have them ranked by relevance. This means that a more relevant article should get a higher ranking than a less relevant article. Relevance is not an easily definable notion, because results are only relevant or irrelevant to a specific user needing an answer to his query. Besides this, relevance is not a strictly binary concept. An article can be more relevant than another one, even when they both mention the subject the user is looking for. In one article it can be the main subject, while it is only a digression in the other article. In a good ranking, the first article would be ranked higher than the latter. It would be ideal to ask real users to rank the answers to their queries. In most cases this is not feasible, because there are many queries, with even more results. Therefore, some automated way of estimating relevance must be used, and one method that is often chosen is the calculation of similarity between the query and the results. The result that is most similar to the query gets the highest ranking.

The similarity measure that will be used in this research is the cosine similarity. To calculate this, the retrieved documents are represented as vectors in a multidimensional space, where each dimension of the space stands for a term in the dictionary. This dictionary contains all words that appear in the document collection. Using the cosine similarity measure, the similarity between vectors can be calculated. As the name reveals, this measure calculates the cosine of the angle between two vectors. The value

ranges from 0, when the documents have nothing in common, to 1, when the vectors are the same, and the documents too, except for a possible difference in length.
The value of the coordinates in the document vectors can be a simple binary value, 0 if the term is absent in the document, and 1 if the term is present. In the cosine similarity calculation that will be used here, the value is determined by the tf-idf of the term.
The tf-idf value is the product of the tf, term frequency, and the idf, inverse document frequency. The term frequency is a count of how often a term occurs in the document in question. The inverse document frequency is the logarithmic value of the total number of documents, divided by the number of documents that the term appears in. The idf of a rare term is high, and the idf of a frequent term is low. Consequently, rare words appearing often in a single document have a high value in this document's vector. The assumption is that such a rare word is meaningful and topical in the context of the document; the document can be assumed to be at least partly about this word.

2.5. Mean average precision

Now the tools are there to achieve a result to a query, and rank that result set. In order to evaluate the ranked results of a query, some sort of indicator is needed. The Mean Average Precision (MAP) is widely used for this purpose. It is the mean value of the average precision of individual queries. The average precision of a query is calculated by taking the average of all the precision values for the relevant documents at their specific ranks. This precision value is the total number of relevant documents up to that rank divided by the total retrieved documents so far. Take a query has three results, two of which are relevant. These relevant results are ranked 1 and 3. Now the precision at rank 1 is 1/1, precision at rank 2 is not calculated, because that document is not relevant, and the precision at rank 3 is 2/3. So the average precision for this query is (1/1 + 2/3)/2 = 0,83. Following this, the average precision of a query is 1 if all relevant documents are ranked higher than the non-relevant documents.
More formally, as Manning et al. (2008, p.147) explain: "If the set of relevant documents for an information need $q_j \in Q$ is $\{d_1, ..., d_m\}$ and R_{jk} is the set of ranked retrieval results from the top result until you get to document d_k, then

$$\mathrm{MAP}(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} \mathrm{Precision}(R_{jk})$$

"

3. Data

In this chapter the data that is used in the research will be described. There are two datasets: the strikes database of the IISG and the collection of digitised newspaper articles of the Royal Library.

3.1. Strikes

The strikes database of the IISG in the version used for this research contains strikes from 1372 up until December 2006, and consists of 16,427 records. These strikes have been collected by Sjaak van der Velden over the years. An action is considered to be a strike based on three criteria (Van der Velden, 2004, pp. 13-14):

- the action is carried out by *wages-dependent* people, e.g. students' and farmers' actions are not included
- the action involves a *temporary* interruption of work
- the action should be *collective*, which means at least two people should be involved in the strike

Besides strikes, there are also lockouts in the database (683 records). Lockouts are actions where the employer keeps employees from working. These lockouts will not be looked into for this research.

The database is a relational database, consisting of 32 linked tables. For every strike a set of attributes is recorded. The attributes used in this research can be found in Table 1. An example of a strike record as is available online via the IISH website[2] is shown in Figure 1.

One particularly interesting field in the database is the report field. This field is a free text field, and can contain several kinds of information. It often contains a short report or summary of the progress of the strike. If this is the case, the information it contains can be useful in the search for relevant articles. For example, there can be names of negotiators or agitators in this text that are not mentioned in the other attributes of the record. These names may be informative because they are likely to be mentioned in relevant newspaper articles.

[2] http://zoeken.iisg.nl/search/search?action=transform&xsl=strikes-form.xsl&col=strikes&lang=en

Attribute	Value
Company	Zero or more companies where the strike took place
Cities	Zero or more locations where the strike took place
Occupation	Occupation according to HISVAK classification
Demand	Wages, employment conditions, company policy, solidarity, etc.
Sector	Agriculture, industry, manufacturing, trade, etc.
Kind of action	Strike, lockout or other
Strike type	Classical, walkout, relay, etc.
Character	Union, wildcat, unknown
Result	Victory, lost, settled, undecided, unknown, non applicable
Start date	Date
Duration	Number of days
Report	Free text
Numbers	Numbers of strikers, lost workdays.
Trade unions	Zero or more trade unions that were involved

Table 1 Important attributes of a strike

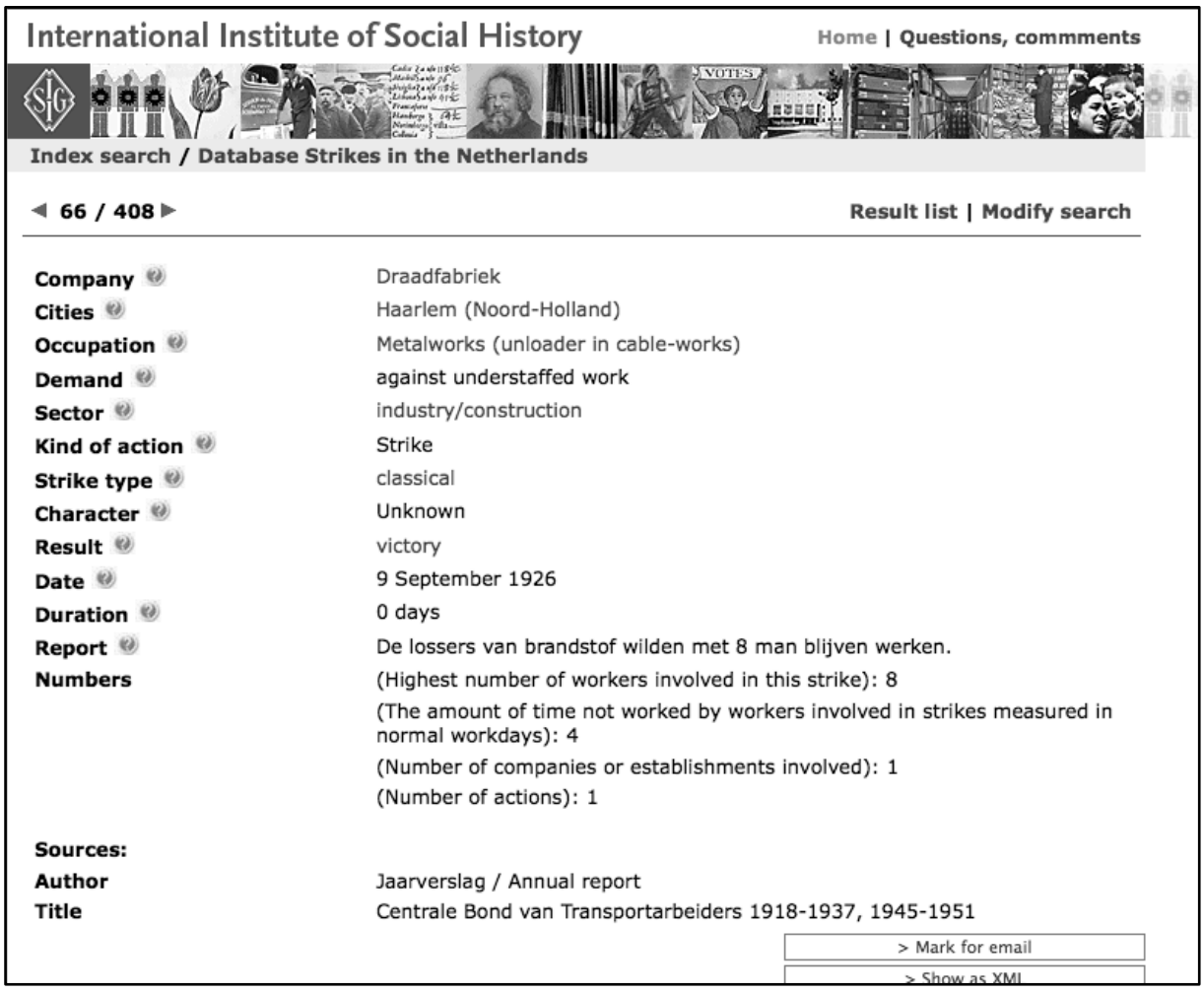

Figure 1 Example of a strike record

In other instances the report field contains some sort of modification of another attribute:

'De duur is een minimum' (The duration is a minimum)

'Kan ook in 1915 geweest zijn' (Could also have been in 1915)

It sometimes contains demands, or results:

'Tegen het ontslag van een collega' (Against the discharge of a colleague)

'Verloren door onderkruiperij' (Lost because of strike-breakers)

Besides the online variant, the strikes database is also made available to the HiTiME project in the form of a Microsoft Access database. In this database the querying possibilities are larger than via the online interfaces.

In this research the focus will be on the strikes that took place from 1910 until 1940. This is because the focus of the HiTiME project is on the period of 1850-1940, and the period within this time frame that is the richest in strikes is 1910-1940. Over half of all recorded strikes happened in this period.

3.2. Newspapers

In 2006 the Royal Library in the Netherlands started the project Databank Digitale Dagbladen[3] (Databank of Daily Digital Newspapers), with the aim to digitise eight million pages of Dutch newspapers by the year 2011. These are newspapers dating from 1618 when the first newspaper was published in the Netherlands, until 1995. Not all newspapers are digitised. A selection has been made based on the importance of a title, with regards to political, cultural and religious, social, economical and journalistic criteria. Only 8% of the total of the 8000 newspaper titles available will be digitised (Klijn, 2009).

In the period of interest to this research, the following titles have been digitised: Algemeen Handelsblad, Amigoe di Curacao, De Sumatra Post, De Tijd, De Waag, Het Centrum, Het Nieuws van den Dag, Het Nieuws van den Dag voor Nederlandsch-Indië, Het Vaderland, Het Volk, Nieuwe Rotterdamsche Courant, Nieuwe Tilburgsche Courant, Rotterdamsch Nieuwsblad, Suriname, and Tilburgsche Courant.

The digitised data is stored in XML-format, using Dublin Core standards. Among the metadata recorded for the newspaper articles are: title, edition, publisher, date, range of distribution, type of article, header of article. For the purposes of this research the most

[3] http://kranten.kb.nl/about/

important metadata are: date, header, and type of the article. The type of the article can be one of four categories: general articles, advertisements, illustrations and births, deaths and marriages.

The digitised newspapers can be accessed in several ways. The access for the general public is via the Historische Kranten website[4]. Here a search can be done on words (using Boolean operators and wildcard characters), date, type of article, titles, range of distribution and place of publication. The articles that form the result set of a query can be viewed in the actual newspaper page (as a pdf-file), or as plain text.

Another way to access the newspapers is via an SRU interface. SRU stands for Search and Retrieval via URLs, and it is a flexible, XML-based protocol. The most common implementation of this protocol is via URL, using the HTTP GET for message transfer (McCallum, 2006 p. 3). The query syntax used with SRU to determine what objects to retrieve is CQL, Contextual Query Language. "CQL is intended to be human readable and writable, and reasonably intuitive, [...] it is more powerful than a simple Google-like language." (McCallum, 2006). A typical query could look like this:

```
http://jsru.kb.nl/sru?query=(staking and haven and dc.type exact
artikel and dc.date>="1924/04/20" and dc.date<="1924/04/30")
&maximumRecords=500
```

Such a query would return an XML file containing records that point to the articles that are the result of the query. This is done via an identifier that is resolved to a URL where the article can be found. An element dc.identifier containing:

```
http://resolver.kb.nl/resolve?urn=ddd:010001322:mpeg21:a0051:ocr
```

is resolved to

```
http://resources2.kb.nl/010000000/articletext/010001322/DDD_010001322
_0051_articletext.xml
```

Thus, the location of the data can be changed, without having to change the metadata referring to it.

One issue that is of importance with all digital data that was not born digital, is the OCR quality. Due to poor printing or deterioration of the original paper, not all characters are recognised for what they really are. The quality of the OCR process varies from page to page in the DDD data. When querying with search words using either the general user interface, or the SRU interface, the search is carried out for the words appearing in exactly that spelling, albeit case insensitive. CQL does have provisions to deal with this,

[4] http://kranten.kb.nl/

such as the relation modifier *fuzzy*, but at the moment this has not been implemented by the Royal Library.

Another issue is the partitioning of the articles. Some large articles consisting of several paragraphs with subheaders have been split up incorrectly into several article files. Also, newspapers in those days had articles titled: *Allerlei* (All-sorts) or *Uit de provincie* (From the province), which contain many news facts rolled into one article. These phenomena are not advantageous to the quality of the information retrieval.

Yet, apart from these issues, the database of newspapers is a great source of real world data.

4. Searching for newspaper articles

In the previous chapters the tools and data used in this research have been described. In this chapter and the next the tools and technologies will be used on the data. This chapter describes a query method that can be used to find newspaper articles that are relevant to particular strikes. After that, a ranking system for the retrieved articles is discussed.

4.1. Querying for articles

The chosen approach to find newspaper articles about strikes, is to devise a general query model that can be adapted for each particular strike. For this a Boolean querying mechanism is used. As seen earlier, this Boolean querying can be processed by the interfaces that the Royal Library has provided for its newspaper database.
Now to build a query model suitable to find most of the relevant articles for any given strike, the strike record provides useful, well-structured information. As with any information retrieval task, a trade-off has to be made between precision and recall. If the query is too precise, relevant articles might be missed. On the other hand, if the query is too broad, non-relevant articles will be retrieved. Given the fact that newspapers report current events, the start date and duration of a strike should be good features. Also the word *staking* (strike) itself should be a good identifier. During the period that is under examination, there were no synonyms used for this phenomenon. In the late nineteenth century, when strikes were starting to happen more often, new terminology had to be established. Foreign terms as *strike* and *grève* were used, besides regional words such as *bollejeije* and *laveij* (Van der Velden, 2004, p. 12). *Staken* and *staking* turned out to be the terms that stuck. Only after the Second World War new terms (and new types of workers' action) were introduced, such as *werkonderbreking*, *stiptheidsactie,* etc. So when looking for articles pertaining to particular strikes, it seems obvious to include this term in the query. The query will be expanded further with data from the strike record, including the industry and occupation names and all the names of persons, places and organisations.

Data from these attributes are used:

- Companies
- Cities (only place names, not province)
- Occupations
- Sectors
- Unions

The report field has been POS-tagged (using Tadpole) and all the proper names are extracted – those are the tokens with the tag SPEC(deeleigen), denoting proper names - to be used as query terms. Only terms that consist of more than two characters are used, to avoid false positives due to OCR-errors in the newspaper articles. Because it is not likely that all the names and terms found in this way will always be present in articles related to a strike, the Boolean query is defined like this:

```
stak?n* AND (term1 OR term2 OR ... OR termn)
date BETWEEN start_date – 7 AND end_date + 3
```

so the word *staking*, or *staken* or *stakend(e(n))* has to be present and at least one of the other terms.

In order to test the effectiveness of the query, strikes were selected from the IISG database that were likely to have appeared in the news. Strikes from 1920 until 1940 were selected, with 250 or more strikers, to ensure that they are large enough for news coverage, and a duration between two and six days, so that the number of articles would not be too large. This resulted in 27 strikes. For these strikes a query was composed as described above and these queries were executed using the public online interface of the KB. All results were checked manually for their relevance to the strike that was queried. An article is deemed relevant if the article is:

- entirely about the strike
- partly about the strike
- a news overview article also mentioning the strike
- about a similar strike in another company concerning the same conflict (this is the case for strike 8738).

The results can be found in Table 2.

ID	Dates	Query terms	Total	Relevant	Precision
5876	9-9-1920 13-9-1920	Groningen, De Arbeid, karrevoerder, slepers	18	4	0.22
6820	1-3-1921 4-3-1921	Gemeente, Onstwedde, Oude Pekela, ontginners landbouwgrond, Scheemda	12	0	0.00
6764	1-5-1921 3-5-1921	Groningen, Groninger Hoogeland, Van der Veen, Nieuw-Amsterdam, dorsers	5	0	0.00
6333	15-5-1922 18-5-1922	HAL, "Müller, Wm. H.", HAR bij Thomsen, Batavierlijn, Batavier IV, Thomsen Lekhaven, Blaauwhoed, Batavier Lijn, Hulllijn, stukgoedwerkers, Rotterdam, SS. Soestdijk, Furness, Thomsen	9	1	0.11
7843	26-5-1922 30-5-1922	HAR, havenwerkers, Centrale Bond, bootwerkers, GEM, Hudig en Veder, Rotterdam	13	9	0.69
8080	11-6-1922 15-6-1922	Houthaven, houtwerkers in de haven, Amsterdam	10	4	0.40
8135	3-10-1922 5-10-1922	HAR, Amsterdam, havenwerkers, bootwerkers	15	6	0.40
8133	1-10-1922 5-10-1922	Schiedam, Nieuwe Waterweg, scheepsbouwers, KCN-organisaties, NVV	4	4	1.00
8183	4-4-1923 6-4-1923	Scheveningen, trawlervissers	2	2	1.00
8289	23-4-1923 25-4-1923	Rotterdam, HAR op Rotterdam-Noord, havenwerkers, bootwerkers	11	8	0.73
8401	9-6-1923 11-6-1923	Zaandam, "Pont, W.", houthandel personeel, NVV	4	3	0.75
8536	17-3-1924 21-3-1924	HAR, Rotterdam, Comite van Actie, Centrale Bond, Algemeen Verkooplokaal, Comite van Tien, havenwerkers, bootwerkers, SVZ, Middelharnis, Katwijk	23	18	0.78
8316	20-8-1924 22-8-1924	VEM, Praxis, Mare$, IJmuiden, zeevissers	7	7	1.00
8621	10-11-1924 12-11-1924	Amsterdam, Rotterdam, havenarbeiders in het overig massagoed, NAS	29	17	0.59
8647	12-1-1925 16-1-1925	Furness, Rotterdam, nageljongen in de scheepsbouw, ANMB	11	0	0.00
8738	22-10-1925 25-10-1925	Wilton, Rotterdam, Schiedam, scheepsbouwers, ANMB	51	40	0.78
8928	28-3-1927 1-4-1927	Valthermond, veenderij, CNV, CAO	1	1	1.00
9305	2-8-1929 6-8-1929	Enka, Hilligerberg, kunstzijde	10	9	0.90
9361	8-10-1929 10-10-1929	HAR, Amsterdam, houtwerkers in de haven, Zaans	15	4	0.27
9240	6-9-1932 9-9-1932	Werkverschaffing, Apeldoorn, Schipbeker	4	1	0.25
10293	15-6-1936 19-6-1936	RDM, Nieuwe Waterweg, Rotterdam, nageljongen in de scheepsbouw, ANMB, Metaalbewerker	13	0	0.00
9182	19-11-1936 21-11-1936	Bendien, Almelo, naaisters	4	4	1.00
10584	6-4-1937 9-4-1937	Blokband, Amsterdam, Taxichauffeur	33	11	0.33
11011	26-4-1937 29-4-1937	Numan's blikfabrieken, Amsterdam, blikwaren, Rijksbemiddelaar	14	1	0.07
9521	2-1-1938 6-1-1938	IJmuiden, zeevissers	2	2	1.00
11352	20-5-1938 22-5-1938	Houthaven, Amsterdam, houtwerkers in de haven, Spijkers	7	2	0.29
11085	9-1-1939 12-1-1939	Amsterdam, HAR aan de houthaven, houtwerkers	6	2	0.33
Total			333	160	0.48

Table 2 Quantity and precision of retrieved newspaper articles

If all the queries are averaged, the precision is 0.48. That is a decent result, given that the query is very broad. Using the ORs, the query is on the recall end of the precision-recall spectrum (see the paragraph on Boolean retrieval). If a more frequent term like Amsterdam is in the query, the precision is likely to drop. Also, the query searches for articles from a week before the start of the strike. This is to include articles mentioning a possible threat of the strike, or ultimatums. However, with spontaneous strikes, taking this week preceding the strike into account will probably lower the precision of the articles retrieved. A check has been made to see if it affected the result if no articles from the week before the strike were collected for strikes with a spontaneous character. This did increase overall precision to 0.56, however, it did cost recall. It turned out that spontaneous does not mean that the strike was only conceived on the day it started, but it means that no unions are involved. For some strikes, such as the harbour strike of June 1922, articles beforehand talk of a call for *'wilde stakingen'* (wildcat strikes). To avoid a loss of information, the choice has been made to aim for recall, and solve the consequential loss of precision in other ways.

4.2. Ranking the results

We now face one of the problems of Boolean retrieval, as Manning et al. (2008) describe: "Boolean queries just retrieve a set of matching documents, but commonly we wish to have an effective method to order (or rank) the returned results." If we have an effective ranking method, we can overcome the problem of the low precision. If the articles can be automatically scored in some way as to how relevant they are to the query, the user will find the most appropriate matches first in the sorted list of results. There are different ways to score retrieved items for a query.

One of the most common scoring systems is based on the amount of similarity between the query and the retrieved items. In order to calculate this similarity, the original strike record is taken as the query document. In this document the following attributes of the strike record are present, including the attribute name (in Dutch): companies, cities, demands, occupations, sector, kind of action, strike type, character, result, date, duration, report, numbers, trade unions. For each retrieved article the similarity between that document and the strike document is calculated, using the cosine similarity measure.

The results of this scoring are evaluated using Mean Average Precision. Only queries of strikes that scored a precision of more than 0 and less than 1 are used in the cosine

similarity calculation, because a ranking of all relevant articles or all non-relevant articles is pointless. So 17 of the original 27 queries will be ranked. The results of the ranking for the queries are shown in Table 3.

ID	Total	Relevant	Precision	Average precision
5876	18	4	0.22	0.16
6333	9	1	0.11	**0.14**
7843	13	9	0.69	0.51
8080	10	4	0.40	0.33
8135	15	6	0.40	0.33
8289	11	8	0.73	0.56
8401	4	3	0.75	**1.00**
8536	23	18	0.78	0.72
8621	29	17	0.59	**0.65**
8738	51	40	0.78	**0.79**
9305	10	9	0.90	0.79
9361	15	4	0.27	0.24
9240	4	1	0.25	**0.50**
10584	33	11	0.33	**0.37**
11011	14	1	0.07	**0.13**
11352	7	2	0.29	**0.31**
11085	6	2	0.33	0.33
Mean			0.46	0.46

Table 3 Results cosine similarity ranking strike articles

The queries where the ranking has improved the precision are shown in boldface. Using a cosine similarity measure has not provided a good ranking, as can be seen from the unranked mean precision (0.46) vs. the ranked MAP (0.46). One reason that this cosine similarity does not result in a good ranking is that the query to which the newspaper articles are compared is not very similar to the articles in the first place. It is, after all, a database record about a strike, and not a newspaper article. This might be improved by augmenting the query documents with words that appear frequently in relevant articles, and not so frequently in non-relevant articles.

To determine those words, a term frequency list has been made of all retrieved relevant articles, and one of all retrieved irrelevant articles. These lists contain the words as tokenised by the Tadpole system, with the total number of times that they appear in their class of documents. From the first list, of the relevant documents, these terms are removed:

- all words consisting of one or two characters
- all names (because they are present in the strike record)

- all words that have a similar frequency for relevant and non-relevant articles, such as stopwords
- all words with a term frequency of less than 15

The remaining list of 84 words is added to each strike query document, and the cosine similarity between this document and the retrieved articles is calculated. The results can be found in Table 4, which contains the same data as Table 3, but with an added column showing the average precision using the extended query documents.

ID	Total	Relevant	Precision	Average precision	Average precision extended query
5876	18	4	0.22	0.16	**0.23**
6333	9	1	0.11	0.14	0.14
7843	13	9	0.69	0.51	**0.52**
8080	10	4	0.40	0.33	0.30
8135	15	6	0.40	0.33	**0.47**
8289	11	8	0.73	0.56	**0.58**
8401	4	3	0.75	1.00	1.00
8536	23	18	0.78	0.72	0.70
8621	29	17	0.59	0.65	**0.69**
8738	51	40	0.78	0.79	**0.82**
9305	10	9	0.90	0.79	0.79
9361	15	4	0.27	0.24	**0.33**
9240	4	1	0.25	0.50	0.50
10584	33	11	0.33	0.37	**0.46**
11011	14	1	0.07	0.13	**0.17**
11352	7	2	0.29	0.31	**0.45**
11085	6	2	0.33	0.33	**0.42**
Mean			0.46	0.46	0.50

Table 4 Results cosine similarity ranking with augmented record

There is an increase in precision for 11 queries (shown in boldface), 4 have the same average precision and 2 have lost average precision. This means that overall, adding the words that are frequent in strike articles has improved the ranking, by a modest 4 percent.

In conclusion, it has proved possible to collect newspaper articles relating to a particular strike with a premodelled query. The precision is not very high, on average just under one in two articles retrieved by the queries is relevant. The ranking using the extended query document is an improvement on the situation without ranking. This means that when the user browses the ranked results, more relevant articles will appear in the first part of the result set. The further down the user scrolls in the set, the lower the percentage of relevant articles will become.

5. Finding strikes that never happened

The previous chapter reported on the retrieval of newspaper articles mentioning particular strikes. In this chapter retrieved articles predating an actual strike are used to extract features, which can be used to find articles that predate counterfactual strikes, the so-called strikes that never happened.

5.1. Strike prelude patterns

In the search for newspaper articles about strikes it has been noticed that sometimes there is already mention of a strike before it has actually occurred. This happens when there is a conflict of some sort between workers and employers, and the workers use the strike as a threat, in order to get their demands met. Using the strike records from the database, it is possible to find these preliminary articles, because then it is clear in which period to look, and the search terms to use. But apart from these conflicts that eventually have resulted in a strike, there must also have been labour disputes where a strike has been avoided for some reason. It might be interesting for historians to find these conflicts and see what differences there are, if any, with the ones that did result in a strike.

To find these counterfactual strikes, we first look for articles preceding strikes that did happen, because we know where to find these. These articles can then be analysed to see what specific features they have. In the IISG strikes database strikes are selected by the following criteria:

```
year BETWEEN 1910 AND 1939
number of strikers >= 500
character = 'Union'
```

The character of the strike is chosen to be Union and not Wildcat or Unknown, because the latter two are not (necessarily) organised in advance.

Using this query on the strikes database results in 108 strikes. With these strikes possible preliminary articles were sought manually in the newspaper database, until a reasonable amount of around a hundred articles was found. The manual search was done with the following query

```
(stak?n* OR term1 OR term2 OR ... OR termn)
date BETWEEN start_date – 30 AND start_date
```

For all the retrieved documents, the relevance was assessed manually. If many articles were found in the first days of the search window an additional search would be done for the month before. Only the articles found relevant were stored. The following strikes yielded a result, as shown in Table 5.

Id	Date	Duration	Terms	No of articles
4267	21-6-1917	10	Mijnwerkers, Limburg	33
5382	22-9-1919	41	Beurtvaarders, landelijk	5
9116	25-5-1928	123	Scheepsbouwers, De Schelde, Vlissingen	22
9255	8-4-1929	180	Zagerij/schaverij, Zaandam	11
10013	16-11-1931	140	Katoenwever, Jannink, Enschede	22
11191	11-5-1938	58	Zeevissers, Katwijk, Scheveningen, Vlaardingen	4
Diverse			Found during other searches	11

Table 5 Strikes used to find preliminary articles

In order to find out what distinguishes these 108 prelude articles from other articles, we can use a machine-learning program such as TiMBL to do a text classification experiment. We need to feed TiMBL with examples of the different classes we have, in this case articles that are about a labour conflict that could lead to a strike, and articles that are not. For the last category one hundred random articles are selected from the period we are investigating.

TiMBL needs features that it can use for distinguishing the categories. For that a frequency list is distilled from the so-called prelude articles. This list is composed of tokens taken from POS-tagged newspaper articles, using Tadpole, just as done earlier with the strike articles. Only content words are taken into account, that means words with tags N (noun), ADJ (adjective) and WW (verb). Stop words such as the verbs *zijn* and *hebben* (*to be* and *to have*) are taken from the list. All words with a frequency of less than five or with a length shorter than three characters are ignored. The reason that such short words are ignored is because chances are very high that OCR errors occur in these words and that false positives or false negatives appear in the search result. There is no comparison made with the frequency of words in the non-prelude articles, because TiMBL should be able to tell which of the words (or features for TiMBL) has a high informative value, and which are less informative.

Now the feature file for TiMBL is created with the 208 words thus found. For each article (prelude and non-prelude) all 208 features get a binary value: 0 if the word does not occur in the article and 1 if the word does occur. TiMBL is then used to calculate the information gain of the different features. In a leave-one-out experiment with TiMBL, using the default IB1 metric, an accuracy of 94.9% was achieved.
Now, TiMBL has rated the 208 features selected (the frequent words in prelude-articles) according to their informative value. The complete list of terms with a gain ratio of more than 0.050 can be found in Appendix A. The top 10 highest scoring terms are shown in the table below.

Gain Ratio	Word
0.324	conflict
0.303	werkgevers
0.298	mijnwerker
0.298	mijnwerkers
0.259	rijksbemiddelaar
0.258	arbeiders
0.254	schelde
0.253	loon
0.240	mijnwerkersbond
0.239	arbeider

Table 6 Top 10 feature words in prelude vs. non-prelude classification

With the terms an extended OR-query is made to search for newspaper articles. All terms with a gain ratio of 0.100 or higher are selected, except for proper names. That means that Schelde, Vlissingen, Jannink, Zaandam, Heerlen and Enschede are not included in the query. This is done because these organisations and locations are very partial to the strikes that were used to collect prelude-articles. This leaves a list of 53 terms that are used to search the articles.

5.2. Finding actual strike threats

Now with this list of words that possibly signal whether an article is about a labour conflict that could lead to a strike, the newspaper database can be searched. The SRU-interface of the newspaper database is used to search for the articles. Using a Perl-program queries are created in the form of URL addresses and the resulting XML-page is

analyzed. For each week during the period 1910-1939 a query is executed using the 53 terms selected earlier connected with ORs, so articles are counted if at least one of the terms is present in the article. As seen earlier, this should lead to a high recall, but low precision. Another query is executed for the same week, without any search terms to determine the total number of articles in that particular week.

The assumption is that when there is a labour conflict going on, there will be more articles scoring positive on the conflict query that week. The percentage of candidate articles for a prelude to a strike varies from 16.1% to 46.1%, so our query has indeed been very broad. The next step is to look for distinct peaks in this percentage. Because the proportion of possible labour conflict rises and falls over time, it is not enough to select the weeks with the highest percentage. Weeks with a high proportion, relative to the surrounding period, should be selected. This can be done by judging the data visually using a graphical representation.

The graphs of three different years are shown in Appendix B. They show a bar for each week, indicating the percentage of candidate prelude articles in the total. The weeks without bars are weeks with a smaller total amount of articles in the database, where the numbers are so few, that the percentage cannot be used as a clear indicator. A week should count at least 250 articles in total to be taken into account. The weeks left out contained in fact just over a hundred or even fewer articles.

Three weeks are selected that could contain a possible labour conflict, one for each decade: 24-30 December 1912, 11-17 September 1926 and 16-22 April 1938. The weeks that are chosen are indicated with an arrow in the graphs.

For these weeks all the articles that scored positive are manually categorised. An article can fall in one of four categories:

1. P (prelude), if the article is about a labour conflict that could lead to a strike. An article is categorised as P, when it speaks of a defined group of workers and employers that are in conflict with each other. The “group” can be a profession, or a company, or an industry. Political discussions that do not mention a specific conflict are categorised as Other.
2. S (strike), if the article mentions an actual strike that is going on at the time of writing, or has ended.
3. F (foreign), if the article is about foreign labour conflicts or strikes.
4. O (other), for all articles that do not fall in one of the above categories.

Some articles can be categorised in more than one category. If this is the case, then the first applicable category is chosen, in the order as they are listed here. The results of this manual assessment for the three weeks chosen is shown in Table 7.

Week	Total	Prelude	Strike	Foreign	Other
24 February 1912	273	12	20	18	223
11 September 1926	133	2	2	2	127
16 April 1938	208	1	1	5	201

Table 7 Results of categorisation possible prelude articles

With the articles categorised like this, the prelude articles have been examined to see if they were or were not followed by an eventual strike.

24 February – 1 March 1912

Twelve prelude articles are found. Some of them mention conflicts that have indeed turned into actual strikes, such as the home workers in the clothing industry, the plasterers, the painters in Groningen, and the journeyman bakers. In two articles there is talk of a possible conflict involving tramcar personnel in The Hague, but this seems to be a conflict still only in the city council, rather than that personnel is yet involved. No other newspaper articles can be found mentioning this unrest, and the first strike of the tramcar personnel is not recorded until 1914.

Three articles mention unrest amongst teachers, about their salaries. No strikes involving teachers happened until 1915, a strike in a specific school, and general strikes about salary in the 1920s. But when the newspapers are consulted about the teachers' unrest in 1912, many articles can be found regarding protests and actions. Of course, during those years there was a strike ban for civil servants in the Netherlands, and it was not until that was lifted in 1974 that the willingness to strike grew for those groups (Van der Velden, 2000, p. 189). The same goes for the group in the last article that possibly concerns a prelude to a strike; this is about the Amsterdam police protesting the long weeks without days off. There is no strike to be found, because even more than teachers, policemen do not strike. The first action that involved policemen took place in 1983, according to the strike database.

11-17 September 1926

Two prelude articles are found, about two different conflicts. One at Lion's butchery in Boxmeer, and for this conflict no strike can be found. The first recorded strike in Boxmeer after this article is in 1946. There is another newspaper article to be found about this conflict, from two weeks earlier.
The other conflict that is mentioned in this week is about transport workers at the Steenkolen Handelsvereeniging (Coal Trading Association) that are disgruntled with the new two-shift system replacing the three-shift system. No related strike can be found in the database, and no other newspaper articles mentioning this conflict are found.

16-22 April 1938

Only one prelude article has been found in this week, regarding a conflict that has been ended without an ultimatum, at Braat company in Delft. Searching the strike database reveals no strikes at this company until 1946. When the newspapers are searched in 1938 for articles about this company, two more articles are found, that date from just before the week under examination, mentioning the threat of a strike at Braat company. So the one prelude article that is found in this week has lead to a strike that never happened.

5.3. Ranking the results

The categorisation of the candidate prelude articles has been all manual work, and it would be very helpful if the articles that are found using the coarse sieve of the Boolean OR-query could be sorted so that the most relevant articles are ranked higher than less relevant ones. We can use similarity metrics to rank the results just as done before with the strike articles. However, in the case of the strike articles, there is a well-defined document to compare the results with, namely the strike record. This record was successfully enhanced with a list of words that appear frequently in strike articles. In the case of the search for prelude articles there is no query to start with. All there is, is the list of words that are frequent in prelude articles, and not in non-prelude articles. We can calculate similarity against this list using the cosine similarity measure. Mean average precision is calculated to see how well this ranking works.
Because only few prelude articles have been found, the average precision has been calculated using only those, but also counting the strike and foreign unrest articles as relevant. The results are shown in Table 8.

Week	Total	Prelude	Precision	Average precision	Prelude, strike and foreign	Precision	Average precision
24 Feb 1912	273	12	0.044	0.116	50	0.183	0.480
11 Sep 1926	133	2	0.015	0.065	6	0.045	0.321
16 Apr 1938	208	1	0.005	0.077	7	0.034	0.489
Mean			0.021	0.086		0.087	0.430

Table 8 Results cosine similarity ranking prelude articles

The ranking using the cosine similarity measure is a great improvement over the not ranked situation.

The search for "strikes that never happened" has resulted in finding some conflicts that never broke out into actual strikes. Each of the three examined weeks yielded result. However, these few results had to be found amongst many retrieved articles that were not relevant to the search question. Using cosine similarity as a ranking tool did a very reasonable job, so that could help in presenting the search results for further analysis.

6. Conclusions and recommendations

In this research two goals were set. The first one was to find newspaper articles related to a particular strike. This has proved to be possible, although the precision has not been too high (0.48 for the 27 strikes the query was tested on). With a ranking system based on the strike record extended with terms that usually appear in newspaper articles about strikes, a Mean Average Precision could be reached of 0.50.
But MAP is calculated over binary relevance scores. The real value of the relevance of the documents may not be so black and white. All documents mentioning the strike in question were deemed relevant, but some of those articles were index type articles, where some article headers were listed, one of them regarding the strike. Other articles have not been correctly split in the Digital Newspaper archive, so two or more articles with different subjects form only one document. To overcome the shortcoming of MAP, that it only works on binary relevance judgements, Kishida (2005) proposes generalised average precision. This new measure does not need binary values, but works with a seven-point sliding scale. A measure like this could be interesting for this research to better evaluate the scoring system.
An argument that could be made against efforts to improve precision, is that of serendipity. This still remains an elusive phenomenon, hard to research because of the importance of the chance factor. However, according to Foster & Ford (2003): "serendipity would appear to be an important concept of the complex phenomenon that is information seeking". If the precision of the query model would become much higher, fewer to no other articles will be included in the result set. These other articles might carry information that is indirectly relevant, because at least some of the queries terms appear in those articles.
Besides precision, there is the issue of recall. This is not measured in this research. It is difficult to measure recall, because all newspaper will have to be spelled out from 1910 until 1940 to be certain that no mention of a strike is overlooked. It would be possible to manually check the newspapers in the date range of the strike query and see if all articles are retrieved. But even then, it is still a labour intensive procedure. If done, it could provide valuable information about how well the query functions, and how it could be improved to include the extra articles found, if any.

One problem that could affect the recall is that spelling and OCR errors are not accounted for. That is an issue that is very relevant when dealing with scanned newspaper articles. Reynaert (2009) has developed a system with which these errors can be automatically detected and corrected. This system, Parallel Text-Induced Corpus Clean-up, or ParTICCL, works with a set of "confusions". These are characters that are confused with others in OCR-ed texts, for instance *in* is often confused with *m*, or *o* with *e*, and vice versa. These confusions are used together with a lexicon of correct words to find all existing erroneous variants of words, and thus be able to propose a correct form. This system could be used to find all variants of a word, and use these all in our queries. For the word *staking* alone, 199 variants can be found in the 1918 volume of Het Volk. These are variants within a Levenshtein distance of 2, so up to two characters in the word are changed. Adding these varieties would expand the queries enormously. It is probably better to wait for the Royal Library to use this system to correct the spelling errors in the articles, as they are working on this. Another option is that the Royal Library makes it possible to use the *fuzzy* search feature. A third option is that the articles are retrieved in the way they are now, but these variants are used when calculating the similarity to achieve a better ranking.

The second goal of the research, trying to find strikes that never happened, has proved to be difficult. There was success in each examined week, but the precision was extremely low. One way this might be improved is by improving the word list that was used to find the articles. This is now based on prelude articles vs. random articles. It could be more appropriate if instead of random articles, articles are sought out that is much closer to the prelude articles. Using these, the border around the class that we are looking for, can be defined much narrower.
There are also things to be said about the way the query terms have been distilled. Yang & Pedersen (1997) have compared different methods of reducing the number of features used for text classification. Their results were that information gain, X^2 and document frequency are good measures to use for that goal. In this research we have used the gain ratio (closely related to information gain, especially for binary features) to reduce the number of features, or rather search terms, we used to query for our articles. This querying can be viewed as a form of text classification. When composing the list of feature words that was fed to TiMBL, term frequency was used to determine which word was used as a feature. It might be an improvement, if document frequency were

used for this, so the number of documents the word appears in, rather than the number of times it appears in the whole set of documents.

And of course, it is up to the historians to determine whether the material we have uncovered, was indeed worth searching for. At least, the search has been a worthwhile experience for me.

7. Bibliography

Daelemans, W., Zavrel, J., Van der Sloot, K. & Van den Bosch, A. (2009). *TiMBL: Tilburg Memory Based Learner, version 6.3, Reference Guide.* ILK Technical Report 10-01. Available from http://ilk.uvt.nl/downloads/pub/papers/ilk1001.pdf

Daelemans, W. & Van den Bosch, A. (2005). *Memory-based language processing.* Cambridge University Press, Cambridge.

Ferguson, N. (1997). *Virtual history: Alternatives and counterfactuals.* Picador, London.

Foster, A. & Ford, N. (2003). Serendipity and information seeking: an empirical study. *Journal of Documentation*, Vol. 59 No. 3, pp. 321-340.

HiTiME project. (2010). Baseline measurement CATCH-HiTiME.

Kishida, K. (2005). *Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments.* Nii technical report (nii-2005-014e), NII.

Klijn, E. (2009). Databank of Digital Daily Newspapers: moving from theory to practice. *News from the IFLA Section on Newspapers,* No. 19, pp. 8-9.

Lee, W.C. & Fox, E.A. (1988) *Experimental comparison of schemes for interpreting Boolean queries.* Technical Report TR-88-27, Computer Science, Virginia Polytechnic Institute and State University.

Manning, C.D., Raghavan, P., Schütze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press, Cambridge.

McCallum, S.H. (2006). *A Look at New Information Retrieval Protocols: SRU, OpenSearch/A9, CQL, and Xquery*. Paper presented at the World Library and Information Congress: 72nd IFLA General Conference and Council, Seoul, Korea.

Reynaert, M. (2009). Parallel identification of the spelling variants in corpora. In *Proceedings of the Third Workshop on Analytics For Noisy Unstructured Text Data*, pp. 77-84.

Salton G., Fox E.A. & Wu H. (1983). Extended Boolean information retrieval. *Communications of the ACM*, Vol. 26 No. 11, pp.1022-1036.

Silver, B.J. 2003. *Forces of Labor. Workers' Movements and Globalization since 1870.* Cambridge University Press, New York.

Van den Bosch, A., Busser, G.J., Daelemans, W., & Canisius, S. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch, In F. van Eynde, P. Dirix, I. Schuurman, and V. Vandeghinste (Eds.), *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, Leuven, Belgium, pp. 99-114.

Van der Velden, S. (2000). *Stakingen in Nederland. Arbeidersstrijd 1830-1995*. Stichting Beheer IISG/NIWI, Amsterdam.

Van der Velden, S. (2004). *Werknemers in actie. Twee eeuwen stakingen, bedrijfsbezettingen en andere acties in Nederland.* Aksant, Amsterdam.

Yang, Y. & Pedersen, J.O. (1997) A comparative study on feature selection in text categorization. In *14th International Conference on Machine Learning*, pp. 412–420.

Appendix A Feature words

Table 9 Gain ratio of feature words in prelude vs. non-prelude classification

Gain Ratio	Word	Gain Ratio	Word
0,324	conflict	0,160	organisatie
0,303	werkgevers	0,154	directies
0,298	mijnwerker	0,149	loonen
0,298	mijnwerkers	0,148	dag
0,259	rijksbemiddelaar	0,140	besturen
0,258	arbeiders	0,134	mijnen
0,254	schelde	0,131	schrijven
0,253	loon	0,124	trachten
0,240	mijnwerkersbond	0,114	firma
0,239	arbeider	0,109	n.l.
0,220	minimum	0,108	betaald
0,215	dreigend	0,108	vakbeweging
0,215	vlissingen	0,108	regeling
0,215	mijnstreek	0,107	conferentie
0,210	vergaderingen	0,106	werk
0,210	ultimatum	0,101	enschede
0,210	christelijke	0,101	partijen
0,210	organisaties	0,099	medegedeeld
0,205	textielindustrie	0,097	bond
0,201	staking	0,097	vergadering
0,200	jannink	0,092	verklaard
0,200	invoering	0,081	besloten
0,200	loonsverlaging	0,080	bespreking
0,195	aktie	0,075	volgende
0,194	arbeid	0,074	maandag
0,190	zaandam	0,074	industrie
0,190	minimumloon	0,073	gehouden
0,184	ingewilligd	0,072	bereid
0,184	verklaarden	0,070	kennis
0,184	houwers	0,070	april

0,182	directie
0,179	werktijd
0,173	arbeidstijd
0,173	houtbedrijf
0,173	verdienen
0,173	patroons
0,170	hoofdbestuur
0,167	hoofdbesturen
0,167	konferentie
0,167	partikuliere
0,167	eischen
0,164	heerlen
0,069	akkoord
0,068	verhooging
0,068	zondag
0,068	kolen
0,061	betrokken
0,060	ontvangen
0,057	stellen
0,056	beweging
0,053	voorstellen
0,053	weken
0,050	voorstel

Appendix B Prelude candidate frequencies per week

Figure 2 Percentage of candidate prelude articles per week 1912

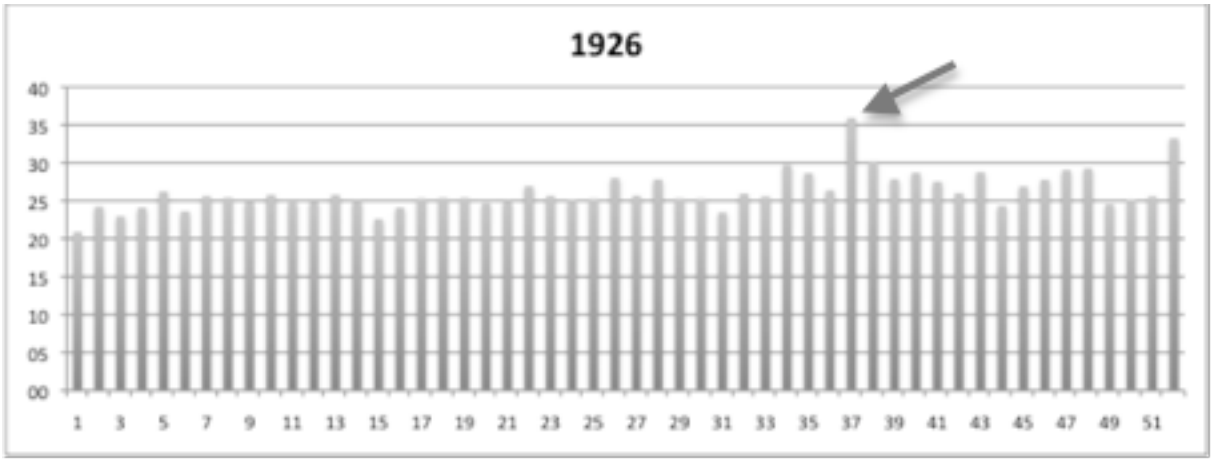

Figure 3 Percentage of candidate prelude articles per week 1926

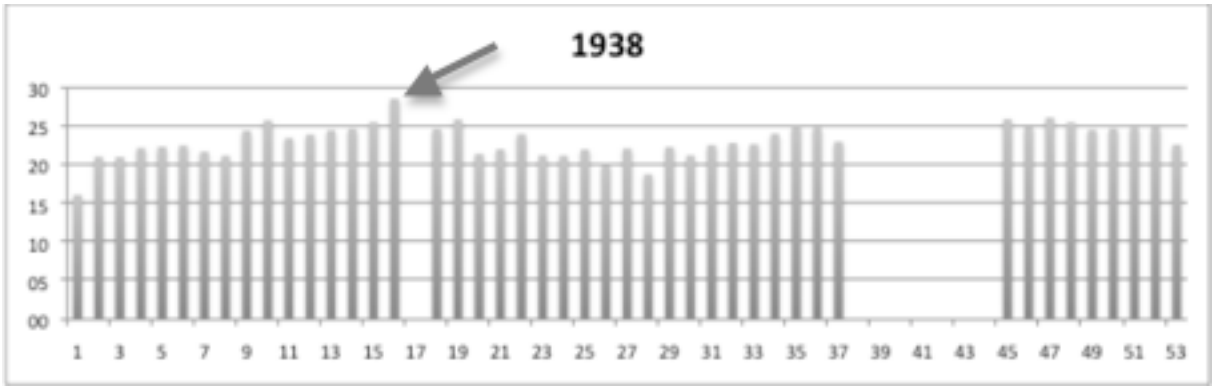

Figure 4 Percentage of candidate prelude articles per week 1938