Beyond Reported History:
Strikes That Never Happened

Martha van den Hoven
TiCC, Tilburg center for
Cognition and Communication
University of Tilburg
Tilburg, The Netherlands
marthov@gmail.com

Antal van den Bosch
TiCC, Tilburg center for
Cognition and Communication
University of Tilburg
Tilburg, The Netherlands
Antal.vdnBosch@uvt.nl

Kalliopi Zervanou
TiCC, Tilburg center for
Cognition and Communication
University of Tilburg
Tilburg, The Netherlands
K.Zervanou@uvt.nl

ABSTRACT
We present a study on applying text analytics methods to historical text and data to uncover aspects of event structure. First, we associate primary historical resources, newspaper articles, to a secondary resource, a database of labor conflicts in the Netherlands, detecting newspaper stories denoting labor conflicts. For the task of retrieving newspaper articles based on database record information, we construct a query model exploiting the database record fields. We consider labor conflicts as historical events referred to in sequences of newspaper article narratives, of which the climax, i.e., the strike, may or may not have occured. We analyse documents preceding a strike by considering them as a sub-narrative class of "strike threat" articles, and we then attempt to retrieve articles referring to conflicts which were about to burst into strike, but for some reason never did: strikes that never happened.

Categories and Subject Descriptors
H.3 [**Information Systems**]: Information Storage and Retrieval; J.5 [**Computer Applications**]: Arts and Humanities

1. INTRODUCTION
The advent of the digital information era has started to expand its effect from transforming conventional information media such as paper archives into digital form, to offering new computational research methods, also to Humanities research areas. In these relatively new domains of application, dedicated research in language and information technologies, for instance the set of tools referred to as text analytics methods, aims to provide solutions for intelligent information storage, access and retrieval.

In historical research, facts and events reported in textual sources play an important role in documenting history. Primary sources of historical information, such as letters, newspapers and brochures, and secondary historical sources, such as biographies and research publications, constitute the principal research material of historical research.

Strikes as an indicator of social and labor unrest are events of interest for social historians. Unfortunately, long-term data collections on events related to labor unrest, such as strikes, exist for very few countries [12]. Much of the current databases are manually compiled by historians, primarily based on investigation of newspaper articles [12, 15]. The development of retrieval and information extraction techniques may contribute by partly automating and speeding up the primary source analysis process. Strikes, if viewed as a series of labour unrest sub-events evolving over a time period and culminating in a strike, may be detected automatically by activity pattens preceding the strike itself. Such an analysis may assist in detecting both conflicts that resulted in a strike, as well as pinpointing conflict periods that were resolved without a strike happening. These resolved conflicts can be characterized as being *counterfactual* [3] strikes, in that the prelude to the non-strike is of the same structure as the prelude to the strike.

In this work, we first attempt to detect associations between secondary historical sources, namely a manually developed strikes database [15] and the respective primary sources; newspaper articles reporting the strikes. Subsequently, we discover patterns of narrated activities preceding a strike, so as not only to automate the detection of eventual strikes, but also to detect conflicts which, for various reasons of potential interest to historians, never resulted in a strike. This work has been carried out within the framework of the HiTiME[1] project [14], a collaboration between the ILK Research Group of Tilburg University[2] and the International Institute of Social History[3] (IISH). The project aims at analysing and associating social history text documents, so as to improve access and to support historical research.

In this paper, we start by a description of our text data and tools used in our method implementation, followed by a presentation and discussion on the two approaches followed in strikes and labour conflict detection: first the association of primary sources, i.e., the articles, to the strikes reported in

---

[1]http://ilk.uvt.nl/hitime/
[2]http://ilk.uvt.nl/
[3]http://www.iisg.nl/

Figure 1: Example of strike record and respective DB fields (online version).

a manually constructed database and, second, the detection of articles narrating the threat of a strike that has never took place.

2. DATA AND METHODS DESCRIPTION

In this section, we describe our research data and the methods used in our research.

2.1 Data

2.1.1 The Strikes Database

The strikes database central to our study has been manually developed for the purposes of historical research [15]. It contains 16,427 records describing strikes, lockouts and other actions, with the earliest record dating from 1372. The majority of the strikes recorded in the database occurred between the years 1900 and 1940. In this work, we only consider strike actions. A *strike* is defined as a specific type of labour action conforming to the following three criteria [16]:

1. It is undertaken by employees only; students' and farmers' actions are not considered;

2. It involves a temporary interruption of work;

3. It is a collective action, involving the participation of at least two persons.

The strikes database has a relational structure, consisting of 32 linked tables. The most important attribute fields for each record are illustrated in Figure 1, which depicts a screenshot of the online Strikes database search interface on the IISH website[4].

One database field of particular importance is the *report* field. As shown in Figure 1, this field contains free text providing information on strikes that could be otherwise classified, such as the names of negotiators, or agitators involved,

---

and, typically, a short report or summary of the strike. Information such as the names of key persons involved is of great use when searching for strike-related articles, as these names are likely to be mentioned in relevant newspaper articles.

In other instances, the report field provides comments on the content of other database fields, such as the strike duration, or the strike date, for example:

'De duur is een minimum'
(The duration is a minumum)
'Kan ook in 1915 geweest zijn'
(Could have also been in 1915)

In other cases, it may refer to the workers demands, or the strike results, for example:

'Tegen het ontslag van een collega'
(Against the dismissal of a colleague)
'Verloren door onderkruiperij'
(Lost because of strikebreakers)

Our research focuses on the strikes reported within the 1910-1940 period, in which over 50% of all recorded strikes in the database occur.

2.1.2 The Daily Digital Newspaper Collection

The Daily Digital Newspaper Collection has been the result of a digitisation project, initiated in 2006, the *Databank Digitale Dagbladen* project[5] (DDD – Databank of Daily Digital Newspapers). This project is undertaken by the Royal Library of the Netherlands and aims at digitizing and providing online access to eight million pages of daily Dutch newspapers by the year 2011. These newspapers constitute a selection of the 8% most important of all newspaper titles, dating from 1618, when the first newspaper was published in the Netherlands, to 1995 [6]. The online DDD website was officially launched on May 27, 2010, making available one million newspaper pages. The digitised data is stored in XML-format, using Dublin Core standards[6]. The collection has been richly annotated with various types of metadata information, the most relevant to our research being the publication date, the headers and the article type. For the latter, a distinction is made between advertisements, illustrations, social announcements (i.e., births, deaths and marriages), and, finally, the news articles which are of interest for our purposes.

The digitised collection may be accessed by the *Historische Kranten* website[7], where a search form interface allows the users to perform queries on both a limited set of metadata, such as date, article type, title, distribution range and publication area, as well as on the newspaper content by boolean and regular expression term queries. The results are available in pdf and plain text. Additionally, the Royal Library provides an alternative mode of access to the collection through an SRU interface. SRU (Search/Retrieval via

---

URL)[8] is a standard XML-focused search protocol for internet search queries which exploits the HTTP GET method for message transfer [9]. The queries are formed in CQL (Contextual Query Language), a query language designed not only to be intuitive and human readable, but also more powerful than languages, such as Google-like languages [9]. In our experiments, we have opted for the SRU interface, because it provides greater flexibility and more querying options. An example of a typical query in CQL for our purposes would be:

```
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
```

This query example would return a single XML file containing pointers (links) to the relevant articles. The pointer to the location of the article is created dynamically by means of article identifier resolution. For example, an element `dc.identifier` containing:

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

would be resolved to the following URL location:

```
http://resources2.kb.nl/010000000/articletext/
010001322/DDD\_010001322\_0051\_articletext.xml
```

This technique allows for the location of the data to be changed, without respective alteration to the metadata referring to it.

Similarly to other digitalised data collections, the DDD collection is affected by OCR quality issues. In particular, poor printing, or deterioration of the original paper, often results in erroneous and inconsistent character recognition. This type of problem affects the retrieval of articles based on search for content words, because this search does not return documents containing spelling variations of the search term. CQL supports fuzzy search, nevertheless this functionality has not yet been implemented by the Royal Library in the current version of the interface.

Another issue affecting the retrieval quality is the article segmentation. In the DDD collection the article segmentation has been performed in a semi-automatic manner, whereby the results of the automatic partitioning were inspected and corrected by humans. We observe however that in some cases large articles consisting of several sections including subheaders have been erroneously split into separate article files, whereas in other cases, newspaper sections containing diverse news stories, such as *Allerlei* (All-sorts), or *Uit de provincie* (From the province), are partitioned as a single article. Unlike the problems related to OCR, the issue of article segmentation is not only the result of current limitations in automatic document segmentation, but it also relates to discourse and document structure issues pertaining to human judgement and collection design decisions.

The problems related to OCR and article segmentation constitute recurrent issues, common to many digitisation efforts and research challenges for language and text analysis researchers.

2.2 Methods Applied

2.2.1 Memory-Based Learning: TiMBL

The task of identifying articles preceding a strike and denoting some form of labour conflict which may, or may not have resulted in a strike action, may be viewed as a boolean text classification learning task, whereby the articles of interest form one class and all other articles, the other. For this purpose, in our approach we have adopted a memory-based learning approach. In particular, we apply the TiMBL[9] implementation of memory-based learning for text classification. TiMBL is a learner which exploits the k-Nearest Neighbour algorithm to perform supervised classification tasks [2, 1]. This algorithm assigns to each new unclassified instance the majority class of its k most similar known instances, i.e., its "k-nearest neighbours".

TiMBL weights each instance features according to the information they reveal about the respective instance class. In our experiments we have used the information gain ratio measure. In estimating information gain, one attempts to measure the informative value of a given feature as the difference between the uncertainty about the class in a situation without knowledge of that feature value, and a situation with that knowledge. A problem with information gain is that it tends to overrate features with many values. The information gain ratio measure alleviates this problem by dividing information gain by the entropy of the feature values [1].

2.2.2 Linguistic Analysis: Tadpole

We exploit the proper noun recognition included in the Tadpole part-of-speech tagger for Dutch, to identify expressions denoting proper names, such as organisations, persons and locations, because these expressions are of particular importance for the retrieval of strike related information. Tadpole[10] is a morphosyntactic tagger and dependency parser for Dutch [13], which relies on TiMBL for its POS tagging and parsing classification tasks. Tadpole initially tokenises the input text and then proceeds to assigning part-of-speech, i.e., grammatical category information, such as noun, verb, etc., to each input text token. Tadpole includes a lemmatiser and a morphological analyser whereby, for a given word surface form, such as *"gesteld"*, its respective affixes are recognised and its lemma *"stellen"* is identified as its canonical, normalised form. Tadpole, finally, includes a dependency parser which creates a graph representing the syntactic dependencies among each sentence constituents.

2.2.3 Boolean and Ranked Retrieval

Boolean retrieval is a classic method of information retrieval. In this type of retrieval, a query term is either found, or not, in a document, thus resulting in the document being considered as relevant, or not, respectively. The query terms may be combined by boolean operators, such as AND, OR,

---

[8]http://www.loc.gov/standards/sru/

[9]http://ilk.uvt.nl/timbl/
[10]http://ilk.uvt.nl/tadpole/

and NOT. Boolean operators, such as AND and OR, affect the retrieval results by applying constraints, or relaxing the matching of query terms to documents [7, 11]. For example, a query searching for '*A and B*' will not return documents merely containing A without B, or B without A, thus limiting the set of results. Conversely, a query searching for '*A or B*' will return both the documents where A and B appear individually, as well as those where A and B are in combination. A query formalism can be extended to support regular expression operators, such as Kleene stars and other character range operators, to allow for fuzzy term matching, i.e., query term matching even in cases where the document term differs slightly by one or more characters from the query term. In other approaches, the retrieval system may use dictionaries so as to allow for matching of normalised, e.g., stemmed words, or lemmas, to surface form word variants, or use a relaxed matching function, such as nearness, where search terms have to be found within a certain distance of each other [8].

The principal problem of the boolean type of retrieval lies in that it provides only limited means to express a graded matching of the documents to a given query. In principle, all document results to a query are treated as equally relevant, although in reality some may be more relevant to a given query than others. Relevance cannot be easily defined in an objective manner, since the results to a query may be relevant or irrelevant depending on the particular user information requirements. Thus, relevance cannot be viewed as a strictly boolean concept. Ranked retrieval approaches provide a solution to these problems by estimating the similarity of a document to a given query and assigning higher ranking to documents closely matched to a query and lower ranking to documents less similar to the query terms. The similarity measure used in this work is the cosine similarity. In order to calculate cosine similarity, documents are represented as vectors in a multidimensional space, where each dimension is a term and the number of dimensions is the total number of terms (i.e., words) appearing in the document collection. Cosine similarity measures similarity as the cosine between such document vectors. The similarity value ranges between 0, when the documents have nothing in common, and 1, when the documents are the same. 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 our implementation of the cosine similarity, this value is determined by the $tf \cdot idf$ of the term. The $tf \cdot idf$ 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.

2.2.4 Evaluation Measures
In this work, we use *Precision* to assess the entire set of our retrieval results and *Mean Average Precision* (MAP) to measure the precision in our ranked retrieval output. Precision is defined as the ratio of the number of correct results divided by the total number of retrieved documents. The Mean Average Precision (MAP) is used to assess the ranking quality of a retrieval method, i.e. whether relevant documents are ranked higher than possibly less relevant, or irrelevant ones, and is estimated as the mean value of the average precision of individual queries. In turn, the average precision of a query is calculated as the sum of all precision values at each rank position of the results. According to Manning et al. [8], if the set of relevant documents for an information need $q_j \epsilon Q$, is $\{d_1, ..., d_m\}$ and R_{jk} is the set of ranked retrieval results from the top result until you get to document dk, then

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

For example, if we have three retrieved documents for a given query, two of which are relevant and ranked at positions 1 and 3 in the ranked results list respectively: 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. Thus, the average precision AP for this query would be:

$$AP = \frac{\frac{1}{1} + \frac{2}{3}}{2} = 0.83$$

In this manner, the MAP of a query result is 1 if all relevant documents are ranked higher than the non-relevant.

3. STRIKE ARTICLES RETRIEVAL
3.1 Boolean Query Modelling
The approach taken to find newspaper articles about strikes is to devise a general query model which can be adapted for each particular strike. For this purpose a boolean querying mechanism is used. As presented in Section 2.1.2, this retrieval method is supported by the Royal Library search interfaces to the newspaper articles collection. For our query model we have to take into consideration the issues discussed in Section 2.2.3, so as to to find the optimal balance between precision and recall by manipulating the boolean query constraints. Our information source for the search query terms is the strikes database records. Our considerations for constructing queries from the database records are as follows.

First, the most general topical term *"staking"* (i.e., strike) itself should be a good identifier. During the period under examination, no synonyms of the word *"staking"* were in use. In the period before our focus period, during the late nineteenth century, when strike activities were becoming common, there was more terminological drift. Foreign loan terms, such as *"strike"* and *"grève"* were used, alongside dialectal terms, such as *"bollejeije"* and *"laveij"* [16] The terms *"staken"* and *"staking"* then became the prevailing terms, at least until after the Second World War, after which new term, such as *"werkonderbreking"* and *"stiptheidsactie"* were introduced to describe new types of labor actions.

Given that newspapers report on current events, the start date and duration of a strike should be important features, as a strike will be a narrative topic for only a limited period, centering around the strike itself. Other salient fields that

our query should include are the affected industry and occupation sector, the location, and the particular persons, workers' unions and companies involved. The *report* field was POS-tagged (using Tadpole) and all tokens are extracted that receive the proper noun tag. These are included in the query terms. In our query term selection, only terms consisting of more than two characters are used, to avoid matching false positives due to OCR-errors in the newspaper articles. The resulting boolean query is formulated as follows:

```
stak?n* AND (term_1 OR term_2 OR ... OR term_n)
date BETWEEN start_date - 7 AND end_date + 3
```

where any form denoting strike (i.e., *staken, stakend, stakenden, staking,* etc.) should be matched alongside any other term originating from the database record for a given time period ranging from a week before the reported start date to three days past the end date. The size of these pre-strike and post-strike windows were estimated by manual inspection of a few cases of strikes, and were judged to cover the majority of articles on specific strikes; selecting higher values would result in lower precision, without likely compensation at the recall side.

In order to test the effectiveness of the query, we select a particular set of strikes from the database likely to be important enough to be reported in the news. We select strikes from the 1920-1940 period, with 250 or more strikers, and a duration between two and six days—longer strikes were avoided to keep the number of query results low to enable manual evaluation, obviously introducing a bias to short strikes. Our selection produces 27 strikes. For each strike a query was composed as described above, and these queries were executed using the web based interface of the Royal Library. All results were checked manually for their relevance to the strike that was queried. An article is deemed relevant if it 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.

Results are illustrated in Table 1. In this table, we show a sample of individual query results, based on the respective database record and indicated by the record ID, namely the total number of retrieved documents, the relevant documents and the respective precision for each query and the entire (All) query results. The MAP for all queries is 48%. In our query formulation, as described above, we have applied broad and relaxed constraints, thus favouring recall rather than precision. For this reason, we consider our precision results satisfactory. Another issue affecting precision is the query terms themselves. For example, when a frequent term such as *"Amsterdam"* is in the query, the precision is likely to drop. Moreover, using the broad pre-strike period aims at retrieving articles mentioning a threat of strike, negotiations, ultimatums, or other signs of unrest. However, in some spontaneous strike cases where the strike was not

Table 1: Strike Articles Sample Retrieval Results

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**
...	...	...	...	...	...
All	333	160	0.48		
Mean			0.46	0.46	0.50

foretold in any newspaper article, this extended time period leads to the retrieval of irrelevant articles. In an experiment we performed on this particular category of strikes, further constrains on the time period increased the overall precision to 56%, while decreasing recall. In order to avoid information loss, aim for higher recall, and attempt to solve the consequential loss of precision in other ways.

3.2 Ranked Retrieval

As discussed in Sec. 2.2.3, a boolean retrieval model considers all document results equally relevant. A ranking method may improve on our average precision if it ranks more relevant documents towards the top. In our implementation of ranked retrieval, we have applied the cosine similarity measure as described in Section 2.2.3. For the similarity estimation, we consider queries as term vectors formed from the content of the database fields, and compare those to the respective document vectors. The ranking results are evaluated using Mean Average Precision measures.

The results are illustrated in Table 1 on the *Average Precision* column. We observe that in some cases indicated by boldface, ranking improves precision. However, in others ranking is ineffective, and the overall MAP is similar to the unranked mean precision, i.e., 46%. We consider that one reason for cosine similarity not resulting in effective ranking is due to the dissimilarity between the articles and the database queries. In order to improve our results, we experimented with query expansion. In particular, we expanded the database record terms in our queries with terms apprearing frequently in relevant articles and not as frequently in irrelevant ones. For this purpose, a term frequency list was compiled for all retrieved relevant articles, and one for all retrieved irrelevant articles. These lists contain terms, as tokenised by Tadpole, alongside their respective frequencies. From the list of relevant articles terms we removed:

- words consisting of one or two characters,
- names (as they are present in the strike record),
- words of similar high frequency in both lists, such as stopwords,

- words with low term frequency (less than 15).

The remaining list of 84 terms was added to each strike query document, and the cosine similarity between this document and the retrieved articles was calculated. The results illustrated in Table 1 on the *Extended Query* column show an increase in average precision in the cases indicated by boldface; the overall mean average precision is improved by 4%.

4. STRIKES THAT NEVER HAPPENED

In order to identify conflicts that never resulted in a strike, we first search and analyse the articles which are known to have lead to a strike. For this purpose we have queried the database for all strikes between 1910-1939 which were organised by a union, as these types tend to be organised and announced in advance:

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

The results were 108 strikes for each of which we identified a preceding article manually, using the following query:

```
(stak?n* OR term_1 OR term_2 OR ... OR term_n)
date BETWEEN start_date - 30 AND start_date
```

Relevance was assessed manually for all retrieved articles. Moreover, if many articles were found in the first days of the search window, an additional search was done for the preceding month. Only articles found relevant were stored.

The task of identifying articles preceding a strike may be viewed as a boolean text classification learning task with a single class, whereby the articles of interest form the positive class and all other articles are negative cases, implicitly. For this task we have used TiMBL, our 108 manually assessed articles as positive examples, and a 100 random articles from the same period as negative. For our article representation model, we have created a frequency list comprising of content words, i.e., those words POS-annotated by Tadpole as nouns, adjectives, or verbs. This list excludes stopwords, forms of the auxiliary verbs *"zijn"* and *"hebben"* (*to be* and *to have*), words of length less than three characters (so as to alleviate OCR errors) and words below frequency of occurrence 5. Contrary to our frequency lists compiled for the *Extended Query* experiment, we do not need to compare frequencies between relevant and non-relevant articles in this case, because TiMBL, as described in Sec. 2.2.1, already weights features according to the information they reveal about the class. The resulting frequency list of 208 words constitutes our classification features. 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 occurs. 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, indicating that articles narrating about

Table 2: Top 10 Gain Ratio scoring features (terms) in 'strike prelude' vs. 'other' articles classification task

Gain Ratio	Term	Gloss
0.324	conflict	*conflict*
0.303	werkgevers	*employers*
0.298	mijnwerker	*miner*
0.298	mijnwerkers	*miners*
0.259	rijksbemiddelaar	*state negotiator*
0.258	arbeiders	*workers*
0.254	Schelde	*Schelde* (name)
0.253	loon	*wages*
0.240	mijnwerkersbond	*mining worker's union*
0.239	arbeider	*workers*

the prelude of a strike can to a reasonable extent be distinguished from other random articles—and that the 208 selected features appear a good starting point for further querying in the full newspaper archive. A sample of our top highest scoring terms is listed in Table 2.

For the formation of our query aimed at finding articles reporting on strike preludes, we selected terms from the set of 208 features with gain ratio $GR \geq 0.100$, and excluded all proper names as those are characteristic of a particular strike. We thus compiled a list of 53 query terms. These terms were subsequently combined in an OR-query to search for newspaper articles for each week of the 1910-1939 period. Another query was performed on each week merely to count the total number of articles in that particular week.

In order to select an appropriate period for our experiments, we assume that on weeks of labour conflict there are more articles matching the conflict query. Our results showed that the percentage of candidate conflict articles in the total pool of articles in a week varies from 16.1% to 46.1%, indicating that our query has been very broad. We identified three weeks in which this percentage of conflict articles distinctly peaked compared to their surrounding period, one for each decade: 24–30 December 1912, 11–17 September 1926, and 16–22 April 1938. For these weeks all the articles that scored positive are manually categorised. An article can be labeled as one of four categories:

P (*prelude*): if the article is about a labour conflict that could lead to a strike. An article is categorised as P, when it refers to a named group of workers and employers in conflict. The group may be a profession, or a company, or an industry. Political discussions not referring to a specific conflict are categorised as *Other*.

S (*strike*): if the article refers to a strike happening at the time of writing, or a stike that has ended.

F (*foreign*): if the article is about foreign labour conflicts such as strikes.

O (*other*): all other articles not belonging to any of the above categories.

In cases where an article may be classified into multiple categories, the first applicable is selected. The results of this

Table 3: Strike Prelude Articles Retrieval Results

Week	Total	Prelude	Precision	Average Precision	Prelude, Strike & 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

manual assessment is shown in Table 3, where *Total* indicates the total number of articles retrieved, *Prelude* the correct prelude articles, and *Prelude, Strike & Foreign* the prelude articles with the other two relevant categories.

To automatically asses our prelude article retrieval, we exploit our ranked retrieval approach to rank our prelude article retrieval results. However, in this case, we do not consider the query terms in estimating similarity; we rather use the entire set of term features characterising prelude articles and calculate document cosine similarity to this term feature list. Results evaluation, similarly to our ranked retrieval evaluation for the database queries task, is based on MAP. As illustrated in Table 3, we have estimated MAP for both the relatively low set of strictly prelude articles, (indicated on the left side of the table), as well the MAP taking into consideration the entirety of our correct results indicating conflict, i.e., including strike and foreign article categories (indicated on the right side of the table). Overall, we observe that the results of this automatic process show a great improvement over the unranked precision results for the prelude articles.

5. DISCUSSION

In this work, the first objective 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, and extended with terms that often appear in newspaper articles about strikes, a Mean Average Precision of 0.50 could be reached. In our evaluation, MAP is calculated based on binary relevance. Yet, the actual relevance value of the documents may not be so discrete. All documents mentioning the strike in question were deemed relevant, but some of those articles were overview or index articles in which headlines were listed, one of them regarding the strike. Some 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 which only works on binary relevance judgements, Kishida [5] proposes the generalised average precision measure. This measure does not require binary values; it rather relies on a seven-point sliding scale. The application of a generalised average precision measure could provide a more refined and revealing evaluation of the scoring system, and should be considered in future work, provided that human judges can be recruited to perform the relevance assessment task.

An argument that could be made against efforts to improve precision, is that of serendipity. According to Foster and Ford [4]: *"serendipity would appear to be an important con-*

cept of the complex phenomenon that is information seeking". Striving for maximal precision sometimes bars the end user from discovering information in strictly non-relevant, but still related articles sharing some of the keywords (and thus perhaps part of the topics) that could lead to interesting findings in their own right.

In our evaluation we have not accounted for recall. In order to determine recall for our document collection, we would have to manually check all newspapers from 1910 until 1940, so as to ascertain that no mention of a strike is overlooked. A more feasible alternative would be to manually check only the newspapers corresponding to the respective date range of the strike query. However, this would still be a laborious and time consuming process. If the resources for this task were available, the estimation of recall could provide us interesting information regarding the performance of the query and the ways by which our recall could be improved.

One issue that is expected to affect recall is that of spelling variations and OCR errors, which are unaccounted for in our current approach. Both issues are relevant when dealing with scanned newspaper articles. Reynaert [10] describes a system which automatically detects and corrects OCR errors. This system, *Parallel Text-Induced Corpus Clean-up*, or *ParTICCL*, works with a set of character confusions that frequently occur in OCR-ed texts, for instance *"in"* is often confused with *"m"*, or *"o"* with *"e"*, and vice versa. These confusions are combined with a lexicon of correct words, so as to detect all existing erroneous word variants, and propose the respective canonical form. This system could be used to find all variants of a word, and use these all in our queries. For example, 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, and as the method is reported to be rather accurate [10], this query expansion can be expected to increase recall. As it stands, the Royal Library is implementing this systems in the articles, so that this functionality will come with their search facilities automatically.

The second objective of this research, namely the discovery of strikes that never happened, proved to be challenging, but what the study shows is that automatic shortcuts can be devised that can be expected to offer reasonably fast alternatives for what would otherwise be tediously long manual searches. Our limited case study on three weeks of apparently high labor unrest proved successful, but with low precision, leaving some work to the user, who normally would be an expert user (e.g. a researcher of social history) capable of filtering out the relevant cases.

A potential improvement could apply to the term list used for detecting the respective articles. We based this list on contrasting strike prelude articles with randomly selected other articles. Yet, the border around the class that we are looking for could be better defined, if instead of random articles, we seek articles which are similar to the prelude articles in many aspects, yet are not about the prelude of a strike—they might, for instance, mention the same companies in the context of events unrelated to labor unrest. Another potential improvement could result from the refinement of feature selection used in classification, namely the selection of our query terms. In a comparative assessment of different methods for feature reduction in text classification, Yang and Pedersen [17] find that information gain, χ^2 and document frequency are good measures to use for that goal. In our task, querying can be viewed as a form of text classification. 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, to query for our articles. When compiling the list of feature terms for 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, by determining the number of documents a given word (i.e., term) appears in, rather than its frequency in the entire document collection.

Finally, a qualitative evaluation of our results by expert historians would reveal whether the material we have uncovered was relevant and worth discovering.

6. CONCLUSION

In this work, we have presented a case study that divided into experiments on associating primary historical resources, such as newspaper articles, to secondary resources, such as databases, and, subsequently, experiments on detecting newspaper stories denoting labor unrest. We based our study on a strikes database developed within the context of historical research, and a large collection of digitized newspaper articles from the beginning of the 20th century. We employed two classic retrieval models: boolean and ranked retrieval. The task of detection of newspaper stories denoting labor conflict is viewed in our approach as a single-class classification task. We used for this purpose the TiMBL memory-based approach implementation and its respective information gain ratio functionality to classify our articles and select the most pertinent features for classification. Overall, our results indicate a relatively low precision in both tasks, which can be partly improved by ranking. In our approach we did not optimize precision, as we want to avoid compromising recall, and we assume our retrieval systems to be used by experts who are willing and capable of filtering relevant results also when these constitute less than 5% of the retrieved results—the alternative being to search for these relevant articles in an unlabeled pool of millions of articles.

Future experiments may specifically assess recall and the effects of OCR error to our retrieval. Refinements to our approach may include refinements of feature selection used in classification and fuzzy query matching.

More generally, we aim to connect to work on cross-document summarization, on the narrative of newspaper stories over time, and on real-time tracking of topics in the news, in order to bring hypotheses into our work that would sharpen our definitions of important events in the news, and how to players in these events, as well as the motifs and behavioral patterns in which players can be expected to act in these events.

7. ACKNOWLEDGMENTS

8. REFERENCES

[1] W. Daelemans and A. van den Bosch. *Memory-based language processing.* Cambridge University Press, Cambridge, UK, 2005.

[2] W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch. Timbl: Tilburg memory based learner, version 6.3, reference guide. ILK Technical Report 10-01, University of Tilburg, Tilburg, The Netherlands, 2009.

[3] N. Ferguson. *Virtual history: Alternatives and counterfactuals.* Picador, London, UK, 1997.

[4] A. Foster and N. Ford. Serendipity and information seeking: an empirical study. *Journal of Documentation*, 59(3):321–340, 2003.

[5] K. Kishida. Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments. NII Technical Report nii-2005-014e, National Institute of Informatics, Tokyo, Japan, October 2005.

[6] E. Klijn. Databank of digital daily newspapers: moving from theory to practice. *News from the IFLA Section on Newspapers*, (19):8–9, 2009.

[7] W. Lee and E. Fox. Experimental comparison of schemes for interpreting boolean queries. Technical Report TR-88-27, Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA, 1988.

[8] C. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval.* Cambridge University Press, Cambridge, UK, 2008.

[9] S. McCallum. A look at new information retrieval protocols: SRU, OpenSearch/A9, CQL, and Xquery. In *The World Library and Information Congress: 72nd IFLA General Conference and Council*, Seoul, Korea, 2006.

[10] M. Reynaert. Parallel identification of the spelling variants in corpora. In *Proceedings of the Third Workshop on Analytics For Noisy Unstructured Text Data*, pages 77–84, 2009.

[11] G. Salton, E. Fox, and H. Wu. Extended boolean information retrieval. *Communications of the ACM*, 26(11):1022–1036, 1983.

[12] B. Silver. *Forces of Labor. Workersš Movements and Globalization since 1870.* Cambridge University Press, New York, NY, USA, 2003.

[13] A. van den Bosch, G. Busser, W. Daelemans, and
S. Canisius. An efficient memory-based
morphosyntactic tagger and parser for dutch. In
F. van Eynde, P. Dirix, I. Schuurman, and
V. Vandeghinste, editors, *Selected Papers of the 17th
Computational Linguistics in the Netherlands Meeting*,
pages 99–114, Leuven, Belgium, 2007.

[14] A. van den Bosch, K. Zervanou, M. van de Camp,
M. van den Hoven, S. Hunt, and M. van der Heijden.
Baseline measurement CATCH-HiTiME, version 1.0.
Technical Report HiTiME-2010-1, University of
Tilburg, Tilburg, The Netherlands, May 2010.

[15] S. van der Velden. *Stakingen in Nederland.
Arbeidersstrijd 1830-1995.* Stichting Beheer
IISG/NIWI, Amsterdam, The Netherlands, 2000.

[16] S. van der Velden. *Werknemers in actie. Twee eeuwen
stakingen, bedrijfsbezettingen en andere acties in
Nederland.* Aksant, Amsterdam, The Netherlands,
2004.

[17] Y. Yang and J. Pedersen. A comparative study on
feature selection in text categorization. In *Proceedings
of the 14th International Conference on Machine
Learning*, pages 412–420, 1997.