First Results on Detecting Term Evolutions*

Nina Tahmasebi
L3S Research Center
Appelstr. 9a
Hannover, Germany
tahmasebi@L3S.de

Sukriti Ramesh
L3S Research Center
Appelstr. 9a
Hannover, Germany
ramesh@L3S.de

Thomas Risse
L3S Research Center
Appelstr. 9a
Hannover, Germany
risse@L3S.de

ABSTRACT

The archival of content like publications or web pages is just the first step toward "full" content preservation. It also has to be guaranteed that content can be found and interpreted in the long run. The correspondence between the terminology used for querying and the one used in content objects to be retrieved, is a crucial prerequisite for effective retrieval technology. However, as terminology evolves over time, a growing gap opens between older documents in (long-term) archives and the active language used for querying such archives. Thus, technologies for detecting and systematically handling terminology evolution are required to ensure "semantic" accessibility of archived content in the long run. The core of our approach is to derive mappings between terminologies originating from different times by the fusion of term concept graphs. To verify the suitability of our approach, we present first results of experiments conducted on The Times archive that covers 200 years of documents. In addition, we discuss how our approach can be applied to web archives and the challenges that arise from this.

Categories and Subject Descriptors

H.3.6 [**Library Automation**]: Large text archives; H.3.1 [**Content Analysis and Indexing**]: Linguistic processing

General Terms

Terminology Evolution, Semantics, Information Extraction

1. INTRODUCTION

Preserving knowledge for future generations is a major reason for collecting all kinds of publications, web pages, etc. in archives. However, ensuring the archival of content is just the first step toward "full" content preservation. It also has to be guaranteed that content can be found and interpreted in the long run.

*This work is partly funded by the European Commission under LiWA (IST 216267)

This type of *semantic* accessibility of content suffers due to changes in language over time, especially if we consider time frames beyond ten years. Language changes are triggered by various factors including new insights, political and cultural trends, new legal requirements, high-impact events, etc. Due to this terminology development over time, searches with standard information retrieval techniques, using current language or terminology would not be able to find all relevant content created in the past, when other terms were used to express the same sought content. To keep archives semantically accessible it is necessary to develop methods for automatically dealing with terminology evolution.

The act of automatically detecting terminology evolution given a corpus can be divided into two subtasks. The first one is to automatically determine, from a large digital corpus, the senses of terms. This task is generally referred to as word sense discrimination.

Word sense discrimination should be differed from word sense disambiguation. The task of word sense disambiguation is, given an occurrence of an ambiguous word and its context (usually sentence or surrounding words in a window), to determine which sense is referred to. Usually the senses used in word sense disambiguation come from explicit knowledge banks such as thesauri or ontologies. Word sense discrimination, on the other hand, is the task of automatically finding the senses of words present in a collection. If an explicit data bank is not used, word sense discrimination can be considered a subtask of word sense disambiguation. Using word sense discrimination instead of a thesaurus or other explicit databanks has its advantages. The method can be applied to domain specific corpora where few or no knowledge banks can be found. These domains could be detailed technical data such as biology or chemistry or at the other end of the spectrum, blogs where many slang words or gadget names are used. Due to the time dependency, it is useful to consider methods that start from unannotated text alone without using existing dictionaries as they do not reflect the time dependency adequately. The output of the first step is a set of clusters representing word senses present in each collection. The set of clusters associated to one collection constitute a *terminology snapshot*.

The second task involved in detecting terminology evolution takes place once several snapshots have been created using corpora from different periods of time. It is in this step that evolution is detected. First word sense evolution is detected,

where the clusters are tracked over time and from this, term evolution is derived. We use an example to illustrate the steps involved in finding terminology evolution. Consider clusters corresponding to "travel". A cluster from the end of the 18th century would likely contain the terms "horse, carriage, boat, walking" whereas a cluster from the end of the 20th century would contain terms like "plane, train, bus, car, boat, walking". If one would track the evolution of clusters in smaller periods of time, it would be possible to detect that the two clusters are related to the same sense. This would correspond to tracking concepts or word senses over time. To track terminology evolution, more information is needed which can be found using *term concept graphs* defined in Section 3.

The work in this paper is based on our formal problem statement in [16]. Therefore we just give a brief overview of the process and the underlying model and focus on our initial practical solution to verify our general approach, which has to be extended in future work.

The remaining paper is organized as follows. We begin by reviewing related work in Section 2. As a foundation, Section 3 gives an overview about the model and the process and Section 4 presents our initial experiments. In Section 5 we discuss the challenges that the approach has with respect to web archives and finish with conclusions and future work in Section 6.

2. RELATED WORK

We will present related work for the two tasks identified above. For the first task, namely word sense discrimination, we will present state of the art and give a description of the different approaches available. For the task of detecting terminology evolution, to our knowledge, little previous work has been done and thus we mainly investigate state of the art in related areas such as evolution in dynamic networks.

Several methods based on co-occurrence analysis and clustering have been proposed like [4, 13, 14]. Taking semantic structures into account improves the discrimination quality. In Dorow et al. [6, 7] it is shown that co-occurrences of nouns in lists contain valuable information about the meaning of words. A graph is constructed where the nodes are nouns and noun phrases. There exists an edge between two nodes if the corresponding nouns are found separated by "and", "or" or commas in the collection. The graph is clustered based on the clustering coeffient of a node and the resulting clusters contain semantically related terms representing word senses. Another approach of word sense discovery is focusing on pattern discovery, such as the one presented in [4]. In [12] a clustering algorithm called Clustering by Committee is presented. This clustering produces clusters with words that can be considered synonymous. An evaluation method is also proposed, where the discovered word senses can be assessed using WordNet [9].

The output from word sense discrimination is normally a set of terms to describe senses found in the collection. This grouping of terms is derived from clustering and we refer to such an automatically found sense as a cluster. Clustering techniques can be divided into hard and soft clustering algorithms. In hard clustering, an element can only appear in one cluster, while soft clustering allows each element to appear in several clusters. Due to the polysemous property of words, soft clustering is most appropriate for word sense discrimination.

Temporal aspects in information retrieval come in different flavors, such as dealing with temporal information within documents, or with temporally versioned documents, or dealing with temporal evolution of terminologies extracted from documents. According to our analysis not much work has been done on the problem of terminology evolution. Abecker et al. [1] show how medical vocabulary evolved in the MEDLINE system. McCray investigates the evolution of the MESH ontology [2]. In the latter study, psychiatric and psychological terms are manually analyzed and their evolution is studied over 45 years. Terminology evolution can also be observed in other domains. For example, in computer science the Faceted DBLP[1] allows analysis of the evolution of given keywords at different times based on the Semantic GrowBag approach [5]. However, all these approaches assess the evolution manually. Furthermore, the results cannot directly be used by information retrieval systems.

Automatic detection of cluster evolution can aid in automatically detecting terminology evolution.This has been a well studied field in the recent years. One such approach for modeling and tracking cluster transitions is presented in a framework called Monic [15]. In this framework internal as well as external cluster transitions are monitored. The disadvantages of the method are that the algorithm assumes a hard clustering and that each cluster is considered as a set of elements without respect to the links between the elements of the cluster. In a network of lexical co-occurrences, the links can be valuable since the connections between terms give useful information to the sense being presented. In [11], a way to detect evolution is presented which also considers the edge structure among cluster members.

An approach taking into account information from previous collections, FacetNet, is proposed in [8]. FacetNet discovers community structure at a given time step t which is determined both by the observed data at t and by the historic community pattern. FacetNet is unlikely to discover community structure that introduces dramatic evolutions in a very short time period. Depending on the characteristics of the word graph derived from our collections it might be a suitable approach to filter out noise. An alternative method of finding evolutions in networks can be inspired by [10]. Here a method for object identification with temporal dimension is presented. In our setting we could consider each cluster found in a snapshot as one observation of an object. We can then cluster observations from different snapshots in order to determine which senses are likely to belong to the same object and be evolutions of one another. An observation outside of a cluster can be considered similar to the sense represented by the cluster, but not as an evolved version of that sense.

To our knowledge only one previous work has been published in the area of terminology evolution [3]. Using language from the past, the aim here is to find good query re-

[1] `http://dblp.l3s.de/`

formulations of concurrent language. A term from a query can be reformulated with a similar term if the terms in the resulting query are also coherent and popular. Terms are considered similar if they co-occur with similar terms from their respective collections. Our approach advances on this by using word senses to find similar terms rather than pure co-occurrence information. Furthermore our approach gives more advanced knowledge on the evolution such as time information on the valid reformulations.

3. A MODEL FOR TERMINOLOGY EVOLUTION

The problem of automatically detecting terminology evolution can be split into sub problems belonging to the two tasks identified in Section 1. Terminology snapshot creation associated with the first task and merging of terminology snapshots as well as mapping concepts to terms, associated with the second task. The first step is to identify and represent the relation between terms and their intended meanings (concepts) at a given time. We call such a representation a *term concept graph* and a set of these a *terminology snapshot*. Such a snapshot is always based on a given document collection $C^{\delta}_{t_i}$ which is a set of documents D taken from a domain δ in the time interval $[t_{i-1}, t_i]$, where $i = 1, \ldots, N, t_i \in T$.

Terminology Snapshot Creation

Each document $D_j \in C^{\delta}_{t_i}$ contains a set of *terms* $w \in W^{\delta}_{t_i}$. The set $W^{\delta}_{t_i}$ is domain specific and contains all terms ever used in domain δ until time t_i. Since W^{δ} is not known we define the approximation W'^{δ}. At time t_0 the set is empty and $W'^{\delta}_{t_i} = W'^{\delta}_{t_{i-1}} \bigcup terms(C^{\delta}_{t_i})$ for $i = 1, \ldots, N$, where $terms(C^{\delta}_{t_i}) = \{w : \exists D\ w \in D \wedge D \in C^{\delta}_{t_i}\}$.

To represent the relation between terms and their meanings we introduce the notion of *concept* and represent meanings as connections between term and concept nodes in a graph. Let $\mathcal{C}$ be the universe of all concepts. The semantics of a term $w \in W^{\delta}_{t_i}$ is represented by connecting it to its concepts. The edges between terms and concepts inherit the time annotation from the collection on which the terminology snapshot is based. For every term $w \in W^{\delta}_{t_i}$, at least one term concept edge has to exist. We introduce the function ϕ to be a representation of term concept relations as follows

$$\phi : W \times T \rightarrow (W \times \mathcal{P}(\mathcal{C} \times \mathcal{P}(T))) \quad (1)$$

$\mathcal{P}$ denotes a power set, i.e. the set of all subsets. Although ϕ only generates one timestamp for each term concept relation, we introduce the power set already at this point to simplify terminology snapshot fusion. The term concept relations defined by ϕ can be seen as term concept graphs.

Terminology Snapshot Fusion

When we have created several separate terminology snapshots, we want to merge them to detect terminology evolution. A term's meaning has evolved if its concept relations have changed from one snapshot to another or if the concepts it relates to, have changed.

The fusion of two terminology snapshots might be more complicated than a simple graph merging. For example, we might merge two concepts from the source snapshots to a single concept in the target graph. As part of the fusion process we also need to merge the timestamps of the edges. When term and concept are equal in both snapshots, the new annotation is just the union of both source annotations. Thus, we represent the concept relations of a term $w \in W$ as set of pairs $(c_i, \{t_{i_1}, \ldots, t_{i_k}\})$. To shorten the notation we define τ as a set of timestamps t_i, i.e. $\tau \in \mathcal{P}(T)$ and the pairs can be written as (c_i, τ_i). We note that a concept does not have to be continuously related to a term; instead the respective term meaning/usage can lose popularity and gain it again after some time has passed. Therefore, τ_i is not necessarily a set of consecutive timestamps.

We introduce the function ψ which merges two term concept graphs. ψ represents relations between concepts from different snapshots.

$$\psi : (W \times \mathcal{P}(\mathcal{C} \times \tau)) \times (W \times \mathcal{P}(\mathcal{C} \times \tau)) \rightarrow (W \times \mathcal{P}(\mathcal{C} \times \tau)). \quad (2)$$

It should be clear that the set of concepts in the resulting graph of ψ does not necessarily have to be a subset of the set of concepts from the source graphs. ψ can iteratively be applied to a term concept graph from time t_N and the term concept graph containing all knowledge about a term up to time t_{N-1}. The motivation behind using terminology snapshots and term concept graphs is that it allows us to track the evolution of a term by tracking the evolution of its senses.

Mapping Concepts to Terms

The graph resulting from snapshot fusion allows us to identify all concepts which have been related to a given term over time. We cannot directly exploit these relations for information retrieval, but we need to map the concepts back to terms used to express them. To represent this mapping, we introduce the third (and last) of our functions, θ. For a given concept c, $\theta : C \rightarrow \mathcal{P}(W \times \tau)$ returns the set of terms used to express c, together with timestamp sets which denote when the respective terms were in use.

The characteristics of θ are clearly dependent on how we choose to define the merging operation of the concepts in ψ. For example, if two concepts are merged, the term assignment has to reflect this merge.

4. A PRACTICAL APPROACH

For our experiments we use the archive of The Times[2]. The collection consists of approximately 20 million articles from 1785 to 1985. We use a 10% sample by extracting data from 4 consecutive years, every 50 years, to get an overview of the dataset. To create a terminology snapshot for each collection, we start by extracting relevant terms. This is done by first tagging the documents with part-of-speech tags and then extracting nouns. These terms constitute the dictionary from which we build a co-occurrence matrix using grammatical relations such as "and", "or" and commas. The co-occurrence matrix is viewed as a graph and an edge is kept if the corresponding nodes have co-occurred at least 3

[2]http://archive.timesonline.co.uk/tol/archive/

times in the collection. The clustering algorithm presented in [6] is used to cluster the graph. The output of the clustering algorithm will identify the concepts of the snapshot. Two different clustering coefficients are used for clustering. Following [6, 7], 0.5 was used for finding relatively stable senses. Due to a relatively low coverage of the graph, 0.3 was also chosen. This value is less strict in defining senses and hence, gives rise to senses with a higher probability of evolution.

Figure 1 shows some statistics over the proportion of distinct nouns found in the graph compared to the number of clusters. The number of clusters using the stricter coefficient follow well the number of nouns found in the graph. Using the less strict coefficient of 0.3, the trend is still visible but the number of clusters is larger. We see a large increase of clusters found after 1838 compared to earlier. This increase is particularly interesting because it cannot be seen when looking at the total number of distinct words in the collections. The reason for this increase is currently being investigated. The irregular behavior of the graph is due to the filtering process. Currently an edge is kept if the frequency of that edge is at least 3. If the filtering is varied based on the collection size, the appearance of the graph would be smoother.

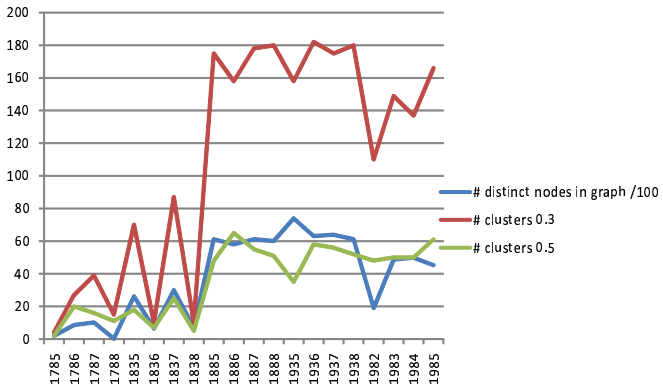

Figure 1: Statistics from a 10% sample of The Times.

We continue by manually analyzing some evolutions found in the collection. Using frequency analysis on the dictionaries, we find that the term "aeroplane", "airplane" and "aircraft" were all used starting from 1935, as can be seen in Fig. 2. While the term "aircraft" was more frequently used during 1935-1985 compared to the other terms, the term "aeroplane" was used most frequently until 1938. "Aeroplane" loses popularity starting from 1982. The term "airplane" is never popular compared to its synonyms, though one could guess that it would be the most frequently used term of the three in a more recent collection. The term "advertisement" and "advertisment" are both present in the collections, Fig. 3. We can draw the conclusion, purely based on frequency and similarity, that "advertisment" is a misspelling of "advertisement".

Both the terms "advertisement" and "aircraft" are found in clusters and show an evolution over the period. The term "advertisement" is clustered together with terms such as "defence, minister, government, argument, marine, pay, fact, country" etc. from the 1935 collection, indicating it to be more propaganda or enlistment ad's than advertisements to sell products. From Fig. 3 we can also see that the word "propaganda" peaks at this period. Later in 1985 the term is clustered with "personality, experience, certificate, career, professional, qualification, training, work, telephone" etc, indicating a profession and hence, is no longer indicative of enlistments for the army.

The term "aircraft" shows similar evolution. While a cluster from 1934 contains terms such as "gun, searchlight, anti" etc, a cluster from 1984 have terms like "transport, training, maintenance, property, vehicle, people, general", etc. Again we see a subtle change in meaning of "aircraft" as it is not seen mainly as a military tool after World War 2.

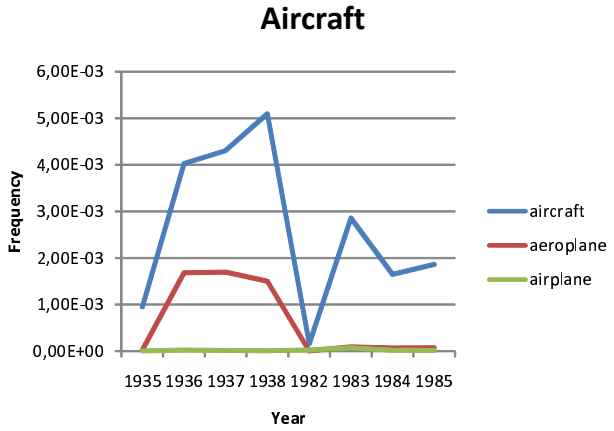

Figure 2: Term frequeny graphs for the terms "airplane", "aeroplane" and "aircraft".

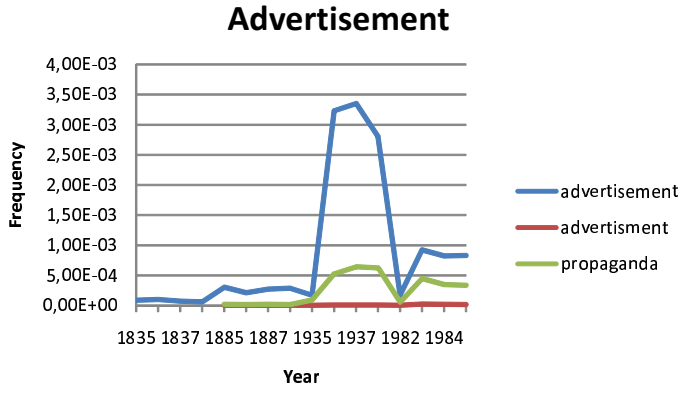

Figure 3: Term frequency graphs for the terms "advertisement", "advertisment" and "propaganda".

Our initial experiments indicate that the clustering algorithm chosen can find clusters corresponding to senses. We see a correspondence of the number of clusters compared to the number of nouns found in the graph. When doing a manual comparison of the clusters, it is clear that the terms

evolve in meaning over time. As discussed above, using frequency analysis can help to see changes, but the clusters give us insight into what types of changes have occurred. The next steps will be to find this evolution in an automatic manner.

5. CHALLENGES IN WEB ARCHIVES

The properties of web archives differ in some aspects from traditional archives of printed documents such as newspaper collections. In traditional archives - even if they are electronic - each document can clearly be distinguished from others as they have a clear time stamp e.g. publishing date or printing date. Furthermore, the content is static and cannot be modified once printed. Due to the static nature of a printed page, it is in the interest of the publisher to publish content of higher quality. Minor mistakes are seldom corrected and ad-hoc publications are an exception. Therefore, the previously described approach for the detection of terminology evolution can be applied directly as the archive property already ensures that each document will be analysed just once. Furthermore, an archive can easily be divided into distinct collections, for example, on a yearly basis, which is necessary to detect evolutions over time.

In contrast to static paper publications, web pages are of a more dynamic nature. Due to their simple structure everybody can publish any content. Once published, the content can easily be changed. Changes range from simple correction of words to entire replacement of the content. The usage of scripting technologies increases the dynamicity of a web page even more. During a single crawl, web archives produce regular snapshots of the web pages, independent of minor or major changes. Between different crawls even unchanged pages are archived more than once.

This has consequences for applying the terminology evolution detection technology to web archives. It is not always possible to determine the real creation time of a page. For example, if content management systems are used, the creation time of a page is often the same as the crawl time. Even if the time information is not reliable, it is the only way to assign a web page to a certain year. To distinguish different versions of a web page the time information is not usable. Therefore different versions of a page have to be distinguished based on their content. As mentioned before the range of changes of two versions of a document can be broad. This raises the next challenge for applying terminology evolution detection technologies to web archives. In case all duplicates or near-duplicates of a page are included in the analysis, the terms of such a page could dominate and therefore bias the results while important new senses would be considered as noise. On the other hand, leaving out duplicates and near-duplicates increases the risk of losing information about senses that are still in use.

A blog site is a typical example of a constantly evolving web site. A blog contains old blog postings as well as new blog postings which are appended. As this page is crawled, old postings as well as new postings become a part of the new collection. Eventually, when the web site has many old postings, the amount of new data will be very small compared to the old postings. This makes it difficult for the system to be able to discover senses found in the new parts of the site, because the older parts are dominating. Eventually the system cannot discover any new senses and hence we cannot find any evolution.

Therefore the terminology evolution method needs to take into account the type of changes to a web page. If just words were corrected or stop words have been changed, then a page should be considered as unchanged. If whole paragraphs are changed within a year these paragraphs should be analysed. If paragraphs change across years, the whole page should be analysed as such a page could contain different terminologies for the same concept. Therefore evolving pages over years could be used as good source for detecting terminology evolution. We will further investigate these issues when we apply our technology to web archives.

6. CONCLUSIONS & FUTURE WORK

Adequately dealing with evolution of terminologies is a necessity to ensure that future generations are still able to access past content even if they are not aware of the changes in the meaning of terms. In this paper we presented our approach to develop an unsupervised method for detecting terminology evolution. The statistics we generated on The Times archive verified the assumption that the evolution of terms and their meanings can be found in an automatic way. Our initial clustering results are promising and will be used for the development of an automatic method to track clusters - and therefore terminology - over time. We furthermore discussed ways to apply our terminology evolution approach to web archives. As web archives differ in their properties from traditional archives, we cannot apply the approach directly in future. However, even if the dynamicity of web pages is an issue, we can make use of it for detecting terminology evolution. This will be part of future work.

7. ACKNOWLEDGEMENTS

We would like to thank Times Newspapers Limited for providing the archive of The Times for our research.

8. REFERENCES

[1] A. Abecker and L. Stojanovic. Ontology evolution: Medline case study. In *Proceedings of Wirtschaftsinformatik 2005: eEconomy, eGovernment, eSociety*, pages 1291–1308, 2005.

[2] Alexa McCray. Taxonomic change as a reflection of progress in a scientific discipline, www.l3s.de/web/upload/talk/mccray-talk.pdf.

[3] K. Berberich, S. Bedathur, M. Sozio, and G. Wiekum. Bridging the terminology gap in web archive search. In *WebDB*, 2009.

[4] D. Davidov and A. Rappoport. Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words. In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 297–304, Sydney, Australia, 2006.

[5] J. Diederich and W. T. Balke. The semantic growbag algorithm: Automatically deriving categorization systems. In *ECDL*, volume 4675 of *Lecture Notes in Computer Science*, pages 1–13. Springer, 2007.

[6] B. Dorow. *A Graph Model for Words and their Meanings*. PhD thesis, University of Stuttgart, March 2003.

[7] B. Dorow and D. Widdows. Discovering corpus-specific word senses. In *EACL '03: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, pages 79–82, Budapest, Hungary, 2003.

[8] Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. L. Tseng. Facetnet: a framework for analyzing communities and their evolutions in dynamic networks. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 685–694, New York, NY, USA, 2008. ACM.

[9] G. A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41, 1995.

[10] S. Oyama, K. Shirasuna, and K. Tanaka. Identification of time-varying objects on the web. In *JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 285–294, New York, NY, USA, 2008. ACM.

[11] G. Palla, A.-L. Barabasi, and T. Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, April 2007.

[12] P. Pantel and D. Lin. Discovering word senses from text. In *In Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 613–619, 2002.

[13] T. Pedersen and R. Bruce. Distinguishing word senses in untagged text. June 09 1997. Comment: 11 pages, latex, uses aclap.sty.

[14] H. Schütze. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123, 1998.

[15] M. Spiliopoulou, I. Ntoutsi, Y. Theodoridis, and R. Schult. Monic: modeling and monitoring cluster transitions. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 706–711, New York, NY, USA, 2006. ACM.

[16] N. Tahmasebi, T. Iofciu, T. Risse, C. Niederée, and W. Siberski. Terminology evolution in web archiving: Open issues. In *8th International Web Archiving Workshop, Aarhus, Denmark, 18th & 19th Sep. 2008*, 2008. `http://iwaw.net/08/IWAW2008-Tahmasebi.pdf`.