The Role of Language Evolution in Digital Archives*

Nina Tahmasebi[1] ** and Thomas Risse[2]

[1] Computer Science & Engineering Department,
Chalmers University of Technology,
412 96 Gothenburg, Sweden
[2] L3S Research Center,
Appelstr. 9, 30167 Hannover, Germany
ninat@chalmers.se, risse@L3S.de

Abstract. With advancements in technology and culture, our language changes. We invent new words, add or change meanings of existing words and change names of existing things. Left untackled, these changes in language create a gap between the language known by users and the language stored in our digital archives. In particular, they affect our possibility to firstly *find* content and secondly *interpret* that content. In this paper we discuss the limitations brought on by language evolution and existing methodology for automatically finding evolution. We discuss measured needed in the near future to ensure semantically accessible digital archives for long-term preservation.

Keywords: language evolution, finding and understanding content, digital archives

1 Introduction

With advancements in technology, culture and through high impact events, our language changes. We invent new words, add or change meanings of existing words and change names of existing things. This results in a dynamic language that keeps up with our needs and provides us the possibility to express ourselves and describe the world around us. The resulting phenomenon is called **language evolution** (or **language change** in linguistics).

For all contemporary use, language evolution is trivial as we are constantly made aware of the changes. At each point in time, we know the most current version of our language and, possibly, some older changes. However, our language does not carry a memory; words, expressions and meanings used in the past are forgotten

* This work is partly funded by the European Commission under ARCOMEM (ICT 270239)

** This work was done while the author was employed at L3S Research Center

over time. Thus, as users, we are limited when we want to find and interpret information about the past from content stored in digital archives.

In the past, published and preserved content were stored in repositories like national libraries and access was simplified with the help of librarians. These experts would read hundreds of books to help students, scholars or interested public to find relevant information expressed using any language, modern or old. Today, because of the easy access to digital content, we are no longer limited to physical hard copies stored in one library. Instead we can aggregate information and resources from any online repository stored at any location. The sheer volume of content prevents librarians to keep up and thus there are no experts to help us to find and interpret information. The same applies to the increasing number of national archives that are being created by libraries which crawl and preserve their national Web. Language in user generated content is more dynamic than language in traditional written media and, thus, is more likely to change over shorter periods of time [TGR12].

Much of our culture and history is documented in the form of written testimony. Today, more and more effort and resources are spent digitizing and making available historical resources that were previously available only as physical hard copies, as well as gathering modern content. However, making the resources available to the users has little value in itself; the broad public cannot fully understand or utilize the content because the language used in the resources has changed, or will change, over time. To fully utilize these efforts, this vast pool of content should be made semantically accessible and interpretable to the public. Modern words should be *translated* into their historical counterparts and words should be represented with their past meanings and senses.

In this paper we will discuss the role of language evolution in digital archives and the problems that arise as a result. We will review state-of-the-art in detecting language evolution and discuss future directions to make digital archives semantically accessible and interpretable, thus ensuring useful archives also for the future. The rest of the paper is organized as follows: In Sec. 2 we discuss different types of evolution and the corresponding problem caused. In Sec. 3 we discuss the differences between digitized, historical content and archives with new content, e.g., Web archives. In Sec. 4 we provide a reivew of current methods for detecting evolution and finally, in Sec. 5 we conclude and discuss future directions.

2 Evolution

There are two major problems that we face when searching for information in long-term archives; firstly *finding* content and secondly, *interpreting* that content. When things, locations and people have different names in the archives than those we are familiar with, we cannot find relevant documents by means of simple string matching techniques. The strings matching the modern name

will not correspond to the strings matching the names stored in the archive. The resulting phenomenon is called **named entity evolution** and can be illustrated with the following:

> "The Germans are brought nearer to Stalingrad and the command of the lower Volga."

The quote was published on July 18, 1942 in The Times [TT42] and refers to the Russian city that often figures in the context of World War II. In reference to World War II people speak of *the city of Stalingrad* or the *Battle of Stalingrad*, however, the city cannot be found on a modern map. In 1961, *Stalingrad* was renamed to *Volgograd* and has since been replaced on maps and in modern resources. Not knowing of this change leads to several problems; 1. knowing only about *Volgograd* means that the history of the city becomes inaccessible because documents that describe its history only contain the name *Stalingrad*. 2. knowing only about *Stalingrad* makes is difficult to find information about the current state and location of the city[3].

The second problem that we face is related to interpretation of content; words and expressions reflect our culture and evolve over time. Without explicit knowledge about the changes we risk placing modern meanings on these expressions which lead to wrong interpretations. This phenomenon is called **word sense evolution** and can be illustrated with the following:

> "Sestini's benefit last night at the Opera-House was overflowing with the fashionable and gay."

The quote was published in April 27, 1787 in The Times [The87]. When read today, the word *gay* will most likely be interpreted as *homosexual*. However, this sense of the word was not introduced until early 20th century and instead, in this context, the word should be interpreted with the sense of *happy*.

Language evolution also occurs in shorter time spans; modern examples of named entity evolution include company names (*Andersen Consulting* $\longrightarrow$ *Accenture*) and Popes (*Jorge Mario Bergoglio* $\longrightarrow$ *Pope Francis*). Modern examples of word sense evolution include words like *Windows* or *surfing* with new meanings in the past decades.

In addition, there are many words and concepts that appear and stay in our vocabulary for a short time period, like *smartphone face*, *cli-fi* and *catfishing* [4] that are examples of words that have not made it into e.g., Oxford English Dictionary, and are unlikely to ever do so.

[3] Similar problems arise due to spelling variations that are not covered here.
[4] http://www.wordspy.com/

2.1 Formal problem definition

Formally, the problems caused by language evolution (illustrated in Figure 1) can be described with the following: Assume a digital archive where each document d_i in the archive is written at some time t_i prior to current time t_{now}. The larger the time gap is between t_i and t_{now}, the more likely it is that current language has experienced evolution compared to the language used in document d_i. For each word w and its intended sense s_w at time t_i in d_i there are two possibilities; 1. The word can still be in use at time t_{now} and 2. The word can be out of use (outdated) at time t_{now}.

Each of the above options opens up a range of possibilities that correspond to different types of language evolution that affect finding and interpreting in digital archives.

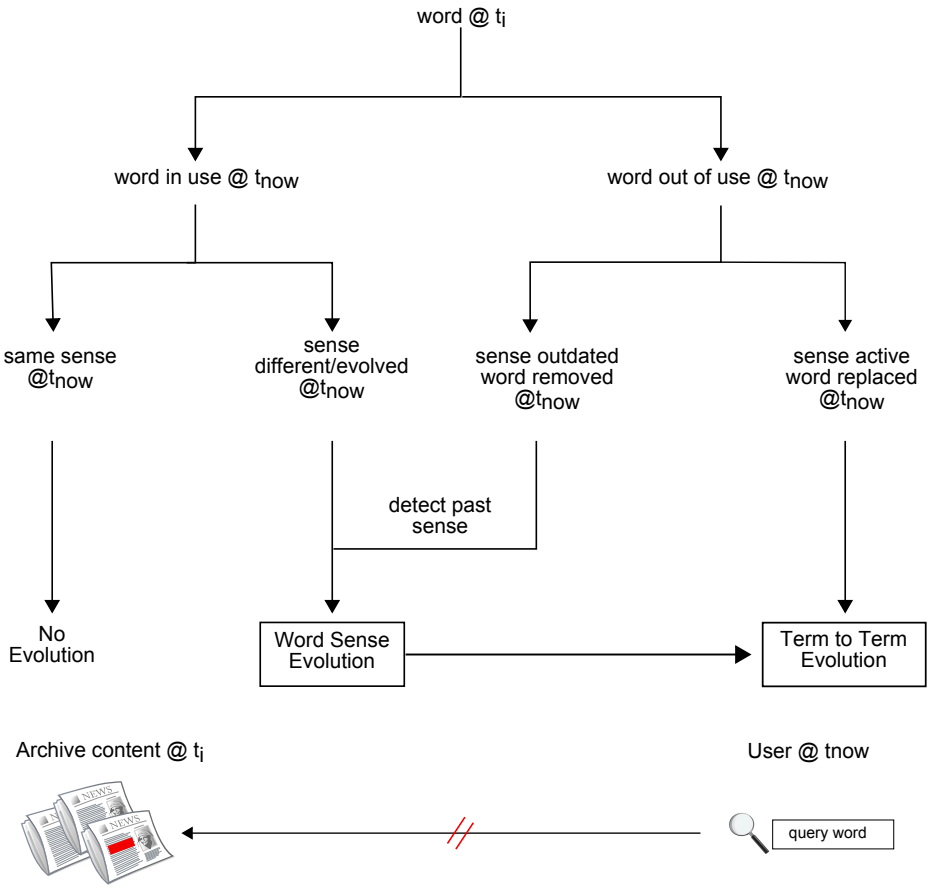

Fig. 1: Diagram of Word Evolution

Word w at time t_i in use at t_{now}

No Evolution: The word is in use at time t_{now} and has the *same sense* s_w and thus there has been no evolution for the word. The word and its sense are stable in the time interval $[t_i, t_{now}]$ and no action is necessary to understand the meaning of the word or to find content.

Word Sense Evolution: The word is still in use at time t_{now} but with a *different sense* s'_w. The meaning of the word has changed, either to a completely new sense or to a sense that can be seen as an evolution of the sense at time t_i. The change occurred at some point in the interval (t_i, t_{now}). We consider this to be the manifestation of word sense evolution.

Word w from t_i out of use at t_{now}

Word Sense Evolution - Outdated Sense: The word is out of use because the word sense is outdated and the word is no longer needed in the language. This can follow as a consequence of, among others, technology, disease or occupations that are no longer present in our society. The word w as well as the associated word sense s_w have become outdated during the interval (t_i, t_{now}). To be able to interpret the word in a document from time t_i it becomes necessary to detect the active sense s_w at time t_i. Because it is necessary to recover a word sense that is not available at time t_{now} we consider this to be a case of word sense evolution.

Term to Term Evolution: The word w is outdated but the sense s_w is still active. Therefore, there must be another word w' with the same sense s_w that has replaced the word w. That means, different words, in this case w and w', are used as a representation for the sense s_w and the shift is made somewhere in the time interval (t_i, t_{now}). We consider this to be term to term evolution where the same sense (or entity) is being represented by two different words. If the word w represents an entity, we consider it to be **named entity evolution**.

In addition to the above types of evolution, there are also *spelling variations* that can affect digital archives; historical variations with different spellings for the same word or modern variations in the form of e.g., abbreviations and symbols. Spelling variations are not considered in this paper.

3 Historical Data vs. Modern Data – Old Content vs. New Content

When working with language evolution from a computational point of view there are two main perspectives available. The first considers today as the point of reference and searches for all types of language evolution that has occurred until today. In this perspective the language that we have today is considered

as common knowledge and understanding past language and knowledge is the primary goal.

In the second perspective the goal is to prepare today's language and knowledge for interpretation in the future. We monitor the language for changes and incrementally map each change to what we know today. We can assume that knowledge banks and language resources are available and all new changes are added to the resources. In the next paragraphs we will discuss the differences between the two perspectives, and the affect on digital archives, in more detail.

3.1 Looking to the Past – The Backward Perspective

When looking to the past we assume that we have the following scenario. A user is accessing a long-term archive and wants to be able to find and interpret information from the past. There are several problems which the user must face. Firstly, there are few or no machine readable dictionaries or other resources like Wikipedia, which sufficiently cover language of the past. The user must rely on his or her own knowledge or search extensively in other resources like encyclopedias or the Web in order to find an appropriate reformulation for modern words. Once the resource is found the user must repeat the process to find the meanings of words, phrases and names in the document. Because of the low coverage of the past, the user can find only limited amount of help in this process.

In order to help users in their research of the past we need to automatically find and handle language evolution. This can be done by making use of existing algorithms and tools or by developing new ones. For both existing and new tools there are severe limitations caused by the lack of digital, high quality, long-term collections. Most existing tools have been designed and trained on modern collections and can have difficulty with problems caused by language evolution. For example, part-of-speech tagging, lemmatization and entity recognition can be affected by the age of the collection and thus limit the accuracy and coverage of language evolution detection which relies on the mentioned technologies.

There is much work being done currently to overcome this lack of resources by digitizing historical documents by means of optical character recognition (OCR). However, many older collections have been stored for a long time which leads to less than perfect quality of the resulting text. Degraded paper, wear or damage as well as old fonts cause errors in the OCR process. This leads to problems in the processing, for example to detect word boundaries or to recognize characters, as well as to verify the results. If words cannot be understood by humans then the correctness of the algorithms cannot be judged. Because of the historical nature of the language, it is also difficult to find people that are qualified to verify, improve or help detect language evolution on such collections.

3.2 Looking to the Future – The Forward Perspective

When looking to the future to find language evolution we have many advantages compared to when looking to the past. The largest advantage is that most resources are born digitally today and thus many of the problems with degraded paper quality and OCR errors are avoided. In addition, there is an abundance of available data. Most concepts, senses and entities are described and referenced over and over again which makes it easier to gather evidence for each one individually.

In addition to the higher amount and quality of the text, there are plenty of tools and resources available that can solve many smaller tasks automatically. Natural language processing tools, machine readable dictionaries, and encyclopedias form an army of resources which can be used to tackle current language. Changes in our world are captured in resources like Wikipedia and questions like *What is the new name of the city XYZ?* can be answered using machine readable resources like Yago [SKW07] or DBpedia [BLK+09]. To prevent information loss in the future, resources like Wikipedia, WordNet and natural language processing tools can be stored alongside the archives. This can significantly simplify finding and verifying language evolution in the future.

Table 1: Processing Comparison - Looking to the Past and Future

Aspect	Past	Future
Content	Digitized after creation, risk of decreased quality.	Increasingly born digital no need for digitization.
Resources	Limited availability	Increasing availability, WordNet, LinkedData etc.
Tools	Mostly modern tools few specialized NLP tools	Existing tools, will be continuously updated
Quality	OCR errors, outdated terms	Noise in user generated text, abbreviations, slang
Crowd sourcing	Limited possibility requires experts	Possible to make use of crowd sourcing

In the perspective of looking to the future we assume that current language is common knowledge and therefore we can employ humans to help detect language evolution. *Crowd sourcing* [How06] is collaborative work performed by large amounts of people and is the mechanism behind creating and maintaining Wikipedia. Such mechanisms could be used to monitor language and detect evolution. If models for representing and storing language evolution are provided,

crowd sourcing could be used to detect language evolution manually or to verify automatically detected language evolution. It is important to note that crowd sourcing is time sensitive and must be done together with the data harvesting to avoid that the crowd forgets.

There are however several limitations. The first limitation is noisy data being published on the Web. With increasing amounts of user generated text and lack of editorial control, there are increasing problems with grammars, misspellings, abbreviations, etc. To which level this can be considered as real noise like with OCR errors is debatable, however, it is clear that this noise reduces the efficiency of tools and algorithms available today. This in turn limits the quality of evolution detection as we depend on existing tools and their efficiency. The second limitation is the restricted nature of resources like Wikipedia. As with dictionaries, Wikipedia does not cover all entities, events and words that exist. Instead, much is left out or only mentioned briefly which limits to which extent we can depend exclusively on these resources.

In order to avoid that future generations face the same problems as we have to face, we need to start thinking about these problems already now. In particular for Web archives that are continuously created and updated, with ephemeral words, expressions and concepts. Otherwise we risk to render a large portion of our archives semantically inaccessible and cannot utilize the great power of crowd sourcing.

4 State-of-the-art

Word Sense Evolution Automatic detection of changes and variations in word senses over time is a topic that is increasingly gaining interest. During the past years researchers have evaluated and researched different parts of the problem mainly in the field of computational linguistics.

[SKC09] presented work on finding *narrowing* and *broadening* of senses over time by applying semantic density analysis. Their work provides indication of semantic change, unfortunately without clues to what has changed but can be used as an initial warning system.

The work presented by [LCM+12] aims to detect word senses that are novel in a later corpus compared to an earlier one and use LDA topics to represent word senses. Overall, the method shows promising results for detecting novel (or outdated) word senses by means of topic modeling. However, alignment of word senses over time or relations between senses is not covered in this work.

[WY11] report on automatic tracking of word senses over time by clustering topics. Change in meaning of a term is assumed to correspond to a change in cluster for the corresponding topic. A few different words are analyzed and there is indication that the method works and can find periods when words change

their primary meaning. In general, the work in this paper is preliminary but with promising indications.

Our previous work presented in [Tah13] was the first to automatically track individual word senses over time to determine changes in the meanings of terms. We found *narrowing* and *broadening* as well as slow shifts in meaning in individual senses and relations between senses over time like *splitting*, *merging*, *polysemy* and *homonymy*. For most of the evaluated terms, the automatically extracted results corresponded well to the expected evolution with regards to the main evolution. However, word senses were not assigned to individual word instances, which is necessary to help users understand individual documents.

In general, word sense disambiguation methods are not sufficient to solve the problem of word sense evolution because discrimination methods 1. often rely on an existing set of word sense; and 2. do not map word senses to each other over time.

Named Entity Evolution Previous work on automatic detection of named entity evolution has been very limited. The interest has largely been from an information retrieval (IR) point of view as named entity evolution makes finding relevant documents more challenging. Unfortunately, no effort has been put towards scalable methods and presentation of evolution to users.

Query reformulation is proposed in [BBSW09] where the degree of relatedness between two terms is measured by comparing co-occurring terms from different time periods. The approach requires recurrent computation for each query as it depends on a target time specified by the user and is not well suited for large datasets.

Semantically identical concepts (nouns) used at different time periods are discovered using association rule mining in [KVB+10]. Entities are associated to events (verbs) and linked across time via the event. The method could be used for shorter time spans but is less suited for longer time spans as verbs are more likely to change over time than nouns [Sag10].

Time-based synonyms (i.e., named entity evolution) are found in [KN10] by utilizing link anchor texts in Wikipedia articles. Unfortunately, link information, such as anchor text, is rarely available in historical archives but might be well suited for Web data.

In our previous work, [Tah13, TGK+12], we proposed NEER, an unsupervised method for named entity evolution recognition independent of external knowledge sources. Using burst detection we find *change periods*, i.e., periods with high likelihood of name change, and search exclusively in these periods for changes. We avoid comparing terms from arbitrary time periods and thus overcome a severe limitation of existing methods; the need to compare co-occurring terms or associated events from different time periods. The method needs to be targeted to Web data and streams of data to avoid re-computation.

In addition to detecting evolution, it is necessary to store evolution and to utilize it for finding and interpreting at query time. Though there is some work done in indexing and retrieval, e.g., [ABBS12, BMRV11], few target the particularities of language evolution.

5 Conclusions and Outlook

Language evolves over time. This leads to a gap between language known to the user and language stored in digital archives. To ensure that content can be found and semantically interpreted in our digital archive, we must consider **semantic preservation** and prepare our archives for future processing and long-term storage. Automatic detection of language evolution is a first step towards offering semantic access, however, several other measures need to be taken. Dictionaries, natural language processing tools and other resources must be stored alongside each archive to help processing in the future. Data structures and indexes that respect temporal evolution are needed to utilize language evolution for searching, browsing and understanding of content. To take full advantage of continuously updated archives that do not require expensive, full re-computation with each update, we must invest effort into transforming our digital archives into **living archives** that continuously learn changes in language.

There are methods for automatically finding language evolution, however, these are initial and have little focus on scalability. Effort needs to be invested into finding large scale methods that provide high quality evolution detection. In addition, the possibility to make use of **crowd sourcing** to improve detection of language evolution should be investigated. Studies are needed to establish where and in which format human input is most beneficial, in particular, when the input is in the form of the crowd without explicit domain expertise. If crowd sourcing solutions are to be employed, the processing must take place at the time of archiving to avoid the crowd forgetting up-to-date changes in the language.

To make the most out of our digital archives, language evolution must be given a **cultural dimension**. For example, the term *travel* has had the same overall meaning over time; *transporting from location A to location B*. However, this does not tell the full story of the word or the concept represented by the word. Today travel is mostly for business or as a happy occasion for holidays, without any substantial risks involved. In the past, traveling contained great dangers and was done at the risk of life. This inherent meaning of a word should be communicated to the user to allow for a full interpretation of language and to entail all dimensions of our language and culture. One possible solution is the **use of images** that can better capture and more easily convey culture.

In addition to viewing language as variant over time, language can be considered variant over demographics. When archiving the Web we have the possibilities to gather knowledge of many subcultures and parts of the world. By continuously

detecting language evolution, we can better determine what content to harvest and store for the future to ensure diverse archives.

Bibliography

[ABBS12] Avishek Anand, Srikanta Bedathur, Klaus Berberich, and Ralf Schenkel. Index maintenance for time-travel text search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 235–244, New York, NY, USA, 2012. ACM.

[BBSW09] Klaus Berberich, Srikanta J. Bedathur, Mauro Sozio, and Gerhard Weikum. Bridging the Terminology Gap in Web Archive Search. In *12th Int. Workshop on the Web and Databases (WebDB'09)*, 2009.

[BLK+09] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia - A crystallization point for the Web of Data. *Journal of Web Semantics*, 7(3):154–165, September 2009.

[BMRV11] Siarhei Bykau, John Mylopoulos, Flavio Rizzolo, and Yannis Velegrakis. Supporting queries spanning across phases of evolving artifacts using steiner forests. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 1649–1658, New York, NY, USA, 2011. ACM.

[How06] Jeff Howe. The Rise of Crowdsourcing. *Wired Magazine*, 14(6), 06 2006.

[KN10] Nattiya Kanhabua and Kjetil Nørvåg. Exploiting time-based synonyms in searching document archives. In *Joint Conference on Digital Libraries (JCDL'10)*, pages 79–88, Australia, 2010.

[KVB+10] Amal Chaminda Kaluarachchi, Aparna S. Varde, Srikanta J. Bedathur, Gerhard Weikum, Jing Peng, and Anna Feldman. Incorporating terminology evolution for query translation in text retrieval with association rules. In *Proceedings of ACM Conf. on Information and Knowledge Management, (CIKM'10), Canada, October 26-30*, pages 1789–1792, 2010.

[LCM+12] Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. Word Sense Induction for Novel Sense Detection. In Walter Daelemans, Mirella Lapata, and Lluís Màrquez, editors, *EACL*, pages 591–601. The Association for Computer Linguistics, 2012.

[Sag10] Eyal Sagi. Nouns are more stable than Verbs: Patterns of semantic change in 19th century English. *The 32nd Annual Conference of the Cognitive Science Society*, 2010.

[SKC09] Eyal Sagi, Stefan Kaufmann, and Brady Clark. Semantic density analysis: comparing word meaning across time and phonetic space. In *Proc. of the Workshop on Geometrical Models of Natural Language Semantics*, GEMS '09, pages 104–111. ACL, 2009.

[SKW07] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA, 2007. ACM.

[Tah13] Nina Tahmasebi. *Models and Algorithms for Automatic Detection of Language Evolution. Towards Finding and Interpreting of Content in Long-Term Archives.* PhD thesis, Leibniz Universität Hannover, To be published 2013.

[TGK+12] Nina Tahmasebi, Gerhard Gossen, Nattiya Kanhabua, Helge Holzmann, and Thomas Risse. NEER: An Unsupervised Method for Named Entity Evolution Recognition. In *Proceedings of COLING 2012*, pages 2553–2568, Mumbai, India, December 2012.

[TGR12] Nina Tahmasebi, Gerhard Gossen, and Thomas Risse. Which Words Do You Remember? Temporal Properties of Language Use in Digital Archives. In *TPDL*, volume 7489, pages 32–37, 2012.

[The87] The Times. Sestini's benefit last night at the Opera-House was overflowing with the fashionable and gay. In *London, England, Apr 27, 1787; pg. 3; Issue 736.* Gale Doc. No.: CS50726043, 1787.

[TT42] DIPLOMATIC CORRESPONDENT The Times. Menace To The Volga. In *London, England, Jul 17, 1942; pg. 3; Issue 49290.* Gale Doc. No.: CS52116209, 1942.

[WY11] Derry Tanti Wijaya and Reyyan Yeniterzi. Understanding semantic change of words over centuries. In *Proceedings of the Int. workshop on DETecting and Exploiting Cultural diversiTy on the social web*, DETECT '11, pages 35–40. ACM, 2011.