Tracking the History of Knowledge Using Historical Editions of Encyclopedia Britannica

M. Gronas[*‡], **A. Rumshisky**[†], **A. Gabrovski**[*], **S. Kovaka**[*], **H. Chen**[*]

MIT, Cambridge, MA, USA[†]
Dartmouth College, Hanover, NH, USA[*]
Center for Media and Society, NES, Moscow[‡]

Abstract

Despite the wealth of newly available digital materials, the scope of text-based investigations has mostly been limited to either synchronous or short-term historical analysis. In this paper, we report on the first stage of the project that focuses on tracking long-range historical change, specifically, on the history of ideas and concepts. The project's aim is to map out the history of representation of knowledge in Europe over last three centuries using as a proxy the history of changes in historical editions of Encyclopedia Britannica. We describe a series of corpus-analytical tasks necessary for building the analytical and comparative tools for historical analysis using scanned noisy text. In this first stage of the project, we focus specifically on the tools for tracking and visualizing the relative importance of people, interconnections between them, and the rise and fall of their reputations.

1. Introduction

Humanities, especially historical disciplines, such as literary history, or political history, or history of science, study changes and evolutions in their respective fields; all such disciplines greatly benefit from tools that map, track down, and measure historical changes and trends within and across their respective fields. But the data in question - e.g. literary and philosophical schools coming and going, scientific theories thriving and falling in disrespect, reputations of authors, musicians, politicians either surviving or falling apart, new gadgets getting invented, new beliefs adopted, fashions, movements, zeitgeists spreading and disappearing, etc. - stems from the pre-digital ages and seems too noisy for digital methods.

In this work, we use historical editions of Encyclopedia Britannica as an (admittedly imperfect) proxy for the history of knowledge in modern Europe. Since its glorious birth in the Age of Enlightenment, the modern encyclopedia has indeed served as the standard representation of human knowledge. The ambition to give both the fullest and most up-to-date account of the state of knowledge gradually transformed the encyclopedia into a never ending collective enterprise, continuously built and rebuilt over centuries by the generations of scholars and editors, our civilization's answer to the Gothic cathedral of the Middle Ages. As the new ideas arise and old ones disappear, fields and domains gain in importance or shrink, encyclopedias respond by mirroring the changes in new editions and updates. Since encyclopedias reflect consensus of scholarly opinions and, by definition, aspire to universality and balance, a single edition may serve as – although obviously imperfect, but best available – representation a synchronic layer of contemporary knowledge, a panoramic snapshot of the state of knowledge. Then the succession of such snapshots, a moving picture of multiple successive editions of the same encyclopedia viewed in a historical perspective, may serve as the best available approximation of the data set on the history of knowledge.

In this paper, we discuss the analytical methods for the historical analysis of conceptual domains. We apply natural language processing and network analysis techniques to the corpus of several historical editions of encyclopedia Britannica; we use the current edition Britannica and to Wikipedia to supplement the analyses. In the present study, we have limited our dataset to the articles about people thus focusing on the social dimension of the history of knowledge. Historical editions in our corpus are OCR scans, and therefore contain very noisy data. In order to use these texts for comparative analysis, we had to perform a series of corpus-analytical tasks, which we describe in this paper. These include:

1. Splitting the text into articles and identifying article titles
2. Identifying articles about people
3. Matching the articles across editions
4. identifying explicit references to other articles.
5. Identifying article categories and matching them across editions.

Once these tasks are performed, the texts can be analyzed for cross-edition changes.

The remainder of this paper is organized as follows. In Section 2., we outline our general approach to formalization of the historical analysis task. In Section 3., we describe the data set we used in our analyses. In Section 4., we describe in some detail the text processing tasks that are required in order to build the diachronic database. We then describe cross-edition analysis methods applied to the derived marked-up text and present the interface for browsing concepts across editions.

2. Formal Description

Our analytical methods are based upon the following generalized model of what a traditional Encyclopedia does:

(1) Selection:
A finite set of areas, concepts, and personalities worthy of inclusion are selected out of the infinity of potential subjects;

(2) Ranking:
The importance of each subject is ranked by the volume assigned to each entry, with more important subject taking up more space;

(3) Interpretation:
Each subject is described and interpreted, and furthermore, placed in relation to other subjects (e.g. through comparison classification, hierarchical subdivisions etc.), thus defining its relative position on the map of contemporary knowledge.

These three steps selection, ranking, and interpretation of the concept and its relative position – inform the changes from one edition to another. Each aspect of concept representation may undergo changes, in particular:

(1) The list of selected subjects grows and changes, picking up new theories, inventions, persons etc, and shedding the ones which are no longer deemed worthy of inclusion;

(2) The volume devoted to a subject also fluctuates mirroring changes in its perceived importance, e.g. the relative space allotted to Shakespeare in Britannica has been steadily growing through 17 and 18 centuries reflecting his growing reputation as the center of the Western Literary Canon, whereas space devoted to, e.g. natural philosophy has been shrinking.

(3) The relations between concepts also evolve through continuous changes in classifications, hierarchies and patterns of associations: in successive editions, a concept may be reclassified, placed in another domain or sub-domain and related to a different set of concepts.

To capture these changes, for each subject (encyclopedic entry on a person) in each of the editions we determine the following relevant factors:

- *Inclusion*, corresponding to presence or absence of an entry in a given edition
- *Size*, corresponding to the percent of the total volume of the edition dedicated to that entry;
- *Centrality*, corresponding to the concept rank, based on several parameters, including the number of incoming references to a given entry, the number of mentions of the subject in other entries, etc.
- *Position*, corresponding to the relations between subjects/domains as represented in concept co-occurrence patterns, clique membership in the graph induced from the set of encyclopedic entries, etc.

The subjects (i.e., in the pilot dataset, persons) are organized into domains, constellations of people belonging to the same field. The historical changes happen both within and across the domains: subjects get included or excluded (factor Inclusion), they grow or shrink in importance (factors size/centrality), they change their position relative to other subjects constituting the domain (factor position). We describe particular tools and methods we develop to track and describe these factors in Section 4..

3. Data set

We used three scanned and OCR'ed out-of-copyright editions of Britannica: Editions 3 and 9 by GoogleBooks, and Edition 11 available from jrank.org. Encyclopedia Britannica granted us research rights to use the electronic text of Britannica's current (15th) edition in XML format. We also used Wikipedia, which effectively provides an (extensive) update to the 15th edition, since the original articles from Britannica's older edition often served as the initial version for Wikipedia entries. These editions gave us the total of 5 points of comparison:

1. Edition 3 (1788–1797)
2. Edition 9 (1875–1889)
3. Edition 11 (1910–1911)
4. Edition 15 (Current EB; 1985–Present)
5. Wikipedia (Present)

Our choice of the specific historical editions was motivated by the fact that some of they represent distinct changes in the state of Encyclopedia Britannica. Edition 3 represents the initial period in the history of Encyclopedia Britannica when it was just being established as an authoritative source. Edition 9, hailed as a Scholar's edition, represented a considerable reworking of the previous editions, with multiple respected authorities in different fields of knowledge contributing articles. Edition 11 again represented a change in the state of knowledge; it was a complete reworking of the Encyclopedia, and remained an authoritative source for several decades.

3.1. Restricting Subject Set

Ideally, the comparative analysis should be conducted using all the subjects for which articles are present in the Encyclopedia, i.e. concepts corresponding to the article titles. This set of concepts should be correlated with and supplemented by the concepts extracted from the text of the articles.

However, even using only the concepts from article titles we encounter a considerable amount of noise. In the present work, we restricted our data to articles on people only for the following reasons. Our analysis is based upon matching of the articles between editions (locating articles on the same topic across edition). Also, in order to conduct the full analysis of conceptual relations and relative importance of different concepts, we detect mentions of different subjects from within other articles, effectively creating a hypertext structure. At present stage, this is exceedingly complicated when dealing with most concepts expressed by common nouns, because of (a) differences in taxonomies: same concept can be part of an article in an earlier edition and has its own article in a later edition b) polysemy: same word can serve as a head word of an article (e.g. nature) and be used in a different meaning (e.g. in proposition by nature). A sturdy disambiguation in such cases is highly problematic. However, proper names (e.g. persons), while still constituting a hugely important domain of knowledge, are less affected by these limitations because : a) persons are usually classified as such; b) namesakes are relatively easy to disambiguate.

4. Tasks

4.1. Cleaning up the data

Historical editions in our corpus are OCR scans, and therefore contain very noisy data. A large proportion of all words are mis-scanned, with text segments from different articles interspersed. The initial task was therefore to (1) split each of the historical editions into separate articles and (2) identify titles for each article. Edition 11 was obtained from the jrank.org website, where it was pre-split into articles. Despite being collectively edited it does contain a significant amount of errors. Edition 11 is also in progress of being manually corrected as part of Project Gutenberg. We replaced the first 13 volumes of text with the manually corrected volumes.

For Editions 3 and 9, we opted not to build a classifier for this auxiliary task. Rather, splitting the text into articles and title identification was performed using a set of simple formatting heuristics, such as looking for uppercase strings at the beginning of paragraphs preceded by blank lines; eliminating mis-scanned tables by identifying text segments with comparatively small average line length, etc. This was complemented by an *alphabetic ordering check* on title candidates. The latter entails checking the alphabetic ordering of the set of proposed titles to remove some of the false positives that break the ordering.

Checking that the next article is in the correct position alphabetically is not sufficient, since one mis-scanned title could cause all subsequent articles to be marked as bad. For example, if there were three articles in a row titled “AARDVARK”, “ABLE”, and “ACE”, they would be all be marked correct because they are in alphabetical order. If, however, “ABLE” was mis-scanned as “AELE” it would still be correct because it still comes after “AARDVARK”, but then “ACE” would be incorrect because it should come before “AELE”. To remedy this, we first find all possible articles by just searching for upper case letters, then assign each article a score based on how many articles before it are in fact alphabetically before, and how many articles after it are in fact alphabetically after, using a threshold to filter out false positives.

Under this setup, a large group of bad titles could cause many nearby articles to get a low score. We therefore first run title extraction a low threshold to get rid most large groups of bad titles, followed by several runs with higher cutoffs to weed out the remaining stragglers. Alphabetic ordering check also had to take into account miscellaneous issues such as the fact in that in some of the older editions letters U and V, as well as J and I, were used interchangeably.

We conducted some accuracy testing by manually checking the accuracy of the split-and-extraction algorithm on a subset of the extracted articles. The following estimates for error rates were obtained:

Edition 3 (GoogleBooks) 19.0% error rate
Edition 9 (GoogleBooks) 10.1% error rate
Edition 11 (Jrank) 14.7% error rate

The OCR’ed editions are quite noisy, and we conducted some quantitative investigations of correctness with a modified spell-checker tool which relies on the current edition of Britannica as well as on Wikipedia for lexical information. For the editions obtained from Google Books, considering only the tokens consisting of alphabetic characters with punctuation, the percentage of misspelled words varied across volumes as follows:

- 7.1–9.6 % in Edition 3
- 5.3-7.9% in Edition 9

4.2. Processing graph structure from individual editions

We have investigated several approaches to constructing article graphs. For each edition, two main types of graphs are currently constructed:

1. Distance graphs, using co-occurrence statistics
2. Explicit reference graphs

We have developed software that allows experimentation with different weighting techniques to tune ranking and clustering methods, with preliminary results available for PageRank and Markov Clustering on both types of graphs for each edition. We ran PageRank on both types of graphs, producing importance ranks for individual articles within each edition.

For each edition, we also ran Markov Clustering on the explicit reference graph in order to partition the articles into distinct clusters. Inflation rate, a factor affecting the segmentation of clusters, was an important part of the the algorithm. This parameter was determined by observing the distance between two different clusterings (the number of node changes required to convert one clustering into another) and cluster tightness. Cluster tightness was determined by the product of the Jaccard similarity to each article’s wikipedia categories and the cluster size. By averaging this score over all clusters for each inflation value, we could objectively select appropriate Markov Clustering parameters for every edition.

The purpose of clustering all articles from every edition is twofold. First, it can serve as a comparison method between articles from within an edition. Second, and more importantly, clusters can track when certain ideas are no longer associated, at least algorithmically, with what it was associated with before.

4.3. Normalization of articles across different editions

We have done cross-edition normalization using Wikipedia categories and the metadata from the current Britannica editions. The cross-edition mapping approach we have been investigating involves the mapping of different categories from the Wikipedia and Current EB to article sets across historical editions. This involves normalization and mapping of article titles to enables the category mapping. Current approaches we are taking involve distributional ranking of article similarity with differential weighting of different article segments, as well as incorporating weighted use of Wikipedia suggestions and targeted Bing searches. We give more detail on this task in the following section.

5. Cross-edition article matching

We used TF*IDF to obtain weighted word vector representations of each article. Since the beginning of the article, i.e. the title and the introduction, usually contain a concise version of the most important information laid out in the rest of the article we over-weighed the beginning of each article, in particular giving more weight to strong identificators such as personal names, dates, names of the professions.

For each edition obtained, every article is first matched to the Wikipedia article on the same topic. The articles matching the same Wikipedia article are then matched across editions as follows.

Step 1. Finding candidates for matching

First we use a list of all article titles in Wikipedia which is sorted alphabetically. An insertion index is obtained for the spot where the title of the article in question can be inserted while preserving the sorted order of the list. Then the surrounding k articles around the insertion index are added to our candidate list (we used $k = 6$ in the experiments below). We then query Wikipedia and Bing for each title of an article and add the top 5 results of each query to our candidate list. The final candidate list of Wikipedia articles is compiled by resolving redirects, removing missing articles and adding candidates from disambiguation pages.

Step 2. Candidate Comparison

Each candidate Wikipedia article is compared to the article we are trying to match using a cosine similarity measure computed for the corresponding weighted TF*IDF word vectors.

Step 3. Detecting articles about people

We apply Wikipedia's categories to filter out non-person articles. Wikipedia articles about people are often assigned categories specifying birth or death year (e.g. the article about Johann Sebastian Bach belongs to the categories 1685 births and 1750 deaths). If an article is not assigned a birth-year or death-year category, the original article from Encyclopedia Britannica is filtered out.

Parameter tuning

The above algorithm uses the following parameters:

1. number of words overweighed in the beginning of article ($first\_word\_lim$);
2. number of words in the beginning considered to the title, usu. names and years ($title\_word\_lim$);
3. weight factor for the title words ($title\_word\_ow$);
4. weight factor for for years found in the beginning of the article ($year\_ow$;
5. weight factor for professions/occupations (e.g. author, poet, tsar) found in the beginning of the article (occ_ow);
6. whether a given word would only be overweighed once; e.g. an occupation might be mentioned multiple times ($only\_once\_ow$)
7. number of words from the article to query Bing and Wikipedia with, along with the title (num_words_q)

We tuned the parameters sing a manually annotated set of 100 articles for each edition. A randomly selected set of articles from Britannica obtained from the initial run of the algorithm was manually matched against the Wikipedia articles and used for parameter tuning. The error rate was then estimated on another 100 of manually matched articles. Table 5. summarizes optimal parameter settings for Edition 9.

The estimates we obtained suggest 75% accuracy, with the error rates for the parameters specified above as follows:

1. incorrect matches - 14.61%
2. non-person articles - 5.26%
3. person articles filtered out - 5.26%

Assigning categories across editions

We used Wikipedia categories to generalize across editions, the rationale is that such a categorization will allow us to track the development of topics, as well as specific articles. Wikipedia's categories, while benefiting from the wisdom of the crowds, also inherit problems associated with it. A lot of categories are ad-hoc and not every article in Wikipedia has been assigned all categories that it should conceptually have. An alternative is to use Encyclopedia Britannica's internal tagging system used in the current electronic edition. Each article in our corpus is matched to its edition 15 counterpart (using our Wikipedia matching as crutch) and get its categories.

6. Browsing Interface and Query Tool

One of the results of our project is an educational online tool for tracking and mapping the social dimension of the history of knowledge, or put simply, a history of reputations. The back-end of the tool is a database that contains all articles about people in different Encyclopedia editions (Britannica 3 ,9 , 11, 15, and Wikipedia). Articles on the same person are matched across editions to compile a master list of people. Each subject is characterized by measures of importance and centrality in respective editions, by the network of co-occurring subjects ("neighbours") , and by the list of categories accompanying this subject in Wikipedia.

The user first picks the domain of interest. The domain is effectively, a Wikipedia category, or a cross-section of Wikipedia categories or lists (e.g. French 19th century composers, chemists, members of the romantic movement, etc). Once the domain is picked, the system produces the snapshots of the domain for each historical edition of the Britannica and for Wikipedia. The snapshots are maps containing all participants of the domain as presented by the respective editions. The more important the subject the larger his or her node on the map; the more central the subject the more central the node; the more frequently two subjects co-occur, the closer they are on the map. The snapshots are then displayed in succession so as to create a movie-like dynamic representation of how domains (actors, their importance, relations between them) changed over the course of last three centuries.

$year\_ow$	occ_ow	$title\_words\_ow$	$once\_only\_ow$	$first\_words\_lim$	$title\_word\_limit$	num_words_q
12	12	8	True	150	4	5

Table 1: Parameter values for Edition 9

6.1. “Gravebook”

One of the products of our research is a novel education tool that highlights and makes more accessible for students the social and intellectual networks of the past. This tool allows one to investigate social and intellectual connections of persons featured in the Current Edition of Britannica and Wikipedia, and reconstruct the underlying social graph, thus creating, effectively, a facebook for the past, or as we term it, Gravebook, an entertaining interface for studying history of human connections. The interface is built as follows. We first determine all entries about people; then all links to other people-articles contained in these entries. For each such link we calculate if the linked person was born after or before the subject of the article; if the life spans of the two overlap, then this connection is considered to be a likely personal acquaintance. Based on this analysis, we will create mock up facebook pages for all persons mentioned in encyclopedias, with linked friends accounts, likes etc. An alternative interface is a 3D visualization of social networks of the past, based on the spring-box algorithms.

7. Related Work

Transmission and diffusion of information, as well as the visualization of connections between different concepts and areas of knowledge, has attracted the attention of scholars is many different research areas. Most of this work has been done in the context of *analysis of social networks*, ranging from product marketing applications (Brown and Reinegen, 1987; Mahajan et al., 1990; Domingos and Richardson, 2001) to the spread of innovation and best practices in medicine and other areas (Coleman et al., 1966; Rogers, 1979) to strategy adoption in game-theoretic settings (Young, 1998; Morris, 2000). Dissemination of ideas in science and the impact of particular works on a given domain has also been the subject of study in scientometrics and bibliometrics, in a long tradition dating back to the works of Rice (Rice, 1965) and Garfield (Garfield, 1955; Garfield, 1972) who pioneered the use of academic *citation patterns* to determine the *impact factor* of the work of particular scientists or scientific journals as publishing entities. The work in scientometrics produced some relevant analytical and visualization tools, e.g. for tracking the life cycle and impact of scientific paradigms (Boyack et al., 2005). An example is the Map of Scientific Paradigms (Boyack et al., 2005) which tracks the relationships between different areas of research in sciences and social sciences by looking at inter-citation and co-citation between scientific journals. Garfield’s work on citation patterns had also given rise to the *link structure analysis* algorithms that have been used more recently in the analysis of web graph structure and the ranking of web pages in search applications, including Google’s *PageRank* (Page et al., 1998), HITS (aka *hubs and authorities* algorithms, or hypertext-induced topic search) (Kleinberg, 1999), and others. More recently, similar methods have been applied to the tracking emergent trends and topics in dynamic data sets, such as email correspondence, news articles, and the blogosphere. For example, Kleinberg (2003)) tracked intensity of topics in email and news articles. Leskovec et al. (2009)) developed the “Memetracker”, a data analysis and visualization tool which tracked the proliferation in the news stories and blog posts of catch phrases mentioned by the candidates in the 2008 U.S. presidential election. Gloor et al. (2008)) measured concept’s relative importance by looking at the number of paths in the network that go through that concept (*betweenness centrality*), applying this method to the documents retrieved from the web for representative phrases in a particular domain, such as names of politicians, brands, etc.

In the analysis of groups in online communities, the focus has been both on (1) identifying subcomminities within a particular network and tracking their development over time (Kumar et al., 2005; Chi et al., 2007), and (2) determining the influential nodes that contribute more to the diffusion of information within the network (Gruhl et al., 2004; Adar and Adamic, 2005; Nakajima et al., 2005; Kimura et al., 2007). The latter has been modeled, for example, on the theory of *propagation of infectious diseases* in epidemiology. For example, Adar and Adamic (2005)) model information “infections” in blogosphere through the analysis of community membership, text similarity, and copying of published URLs between pairs of blogs. How often a link is copied from one blog to another is factored in both determining the influential nodes and in visualization of “infection graphs”.

In the framework of various NLP tasks, such as word sense induction and disambiguation, a wealth of computational methods has been developed for the analysis of distributional similarity between words using word co-occurrence patterns. In such methods, contexts of occurrence for each word, comprised by “bags of words” (Schutze, 1998; Widdows and Dorow, 2002) or features based on grammatical dependencies of a given word are compiled into distributional profiles (Hindle, 1990; Pereira et al., 1993; Grefenstette, 1994; Lin, 1998; Pantel and Lin, 2002; Rumshisky and Grinberg, 2009). Similarity between such profiles is computed using vector space (e.g. cosine), set-theoretic (e.g. Dice, Jaccard), graph-based, or information-theoretic similarity measures (e.g. relative entropy, Jensen-Shannon divergence). While such paradigmatic relations between words are captured through distributional similarity, syntagmatic relations between words and their significant collocates are captured through various association scores (e.g. mutual information, t-test scores) (Church and Hanks, 1990; Kilgarriff et al., 2004). Clearly, changing distributional patterns for words related to a particular concept should be taken into account when modeling change in knowledge representation.

Wikipedia has attracted a lot of research both on the analysis of the resulting concept structures (applying both graph-based and word co-occurrence methods) and the use of the resulting resource in NLP tasks. Effectively, Wikipedia provides two resources, a set of hyperlinked articles (the *article graph*) and a taxonomy-like set of categories with the hyponymy- or meronymy-based subsumption relation (the *category graph*, which is allowed to contain cycles and disconnected nodes). A number of studies in recent literature examined the properties of both graphs, and both graphs have been used to measure semantic relatedness of between concepts and to evaluate the overall semantic structure of covered topics. For example, Zlatić et al. (2006)) examined the link structure of the article graph for such properties as degree distributions, growth, topology, reciprocity, clustering, etc. Zesch and Gurevych (2007)) looked at the applicability of standard semantic relatedness measures to the category graph. Bellomi and Bonato (2005)) applied HITS and PageRank algorithms to the article graph to evaluate the importance of each category. Other parameters have also been used to analyze semantic structure of the *article graph*. Holloway et al. (2007)) used the co-occurrence of categories within individual articles to analyze and and visually map semantic interrelations between categories, comparing the resulting graphs for Wikipedia, Britannica and Microsoft Encarta. Buriol et al. (2006)) have looked at the evolution of the article graph over time using timestamps associated with each article.

8. Conclusions and Future Work

The project has demonstrated the potential for a NLP based analysis of long-range historical datasets. Our research has confirmed the validity of our initial basic hypothesis, namely that the relations between textual entities in an encyclopedia can be used as a proxy for relations between corresponding conceptual entities in the minds of educated contemporaries. The resulting software tools for mapping the history of reputations and social networks of the past have a potential to become useful and entertaining educational tools. In the future we plan to extend our research and tools beyond person articles into the realm of more complicated concepts. This will require additional work on the problem of changing taxonomies and disambiguation. Another important extension of our current research will be an analysis of changes in the relations between different domains, rather than between the members within the same domains.

9. Acknowledgments

This research is sponsored in part by a Digital Humanities Start-Up grant from the National Endowment from the Humanities (HD-51128-10). We would like to thank Britannica for providing us with research access to the current, 15th edition. Historical editions were obtained from GoogleBooks scans. We would like thank Jon Orwant from GoogleBooks for his assistance in obtaining the requisite texts.

10. References

Eytan Adar and Lada A. Adamic. 2005. Tracking information epidemics in blogspace. *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, pages 207–214.

F. Bellomi and R. Bonato. 2005. Network analysis for wikipedia. In *Proceedings of Wikimania*.

K.W. Boyack, R. Klavans, and K. Börner. 2005. Mapping the backbone of science. *Scientometrics*, 64(3):351–374.

J. Brown and P. Reinegen. 1987. Social ties and word-of-mouth referral behavior. *Journal of Consumer Research*, 14(3):350–362.

L.S. Buriol, C. Castillo, D. Donato, S. Leonardi, and S. Millozzi. 2006. Temporal analysis of the wikigraph. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 45–51. IEEE Computer Society Washington, DC, USA.

Yun Chi, Shenghuo Zhu, Xiaodan Song, Junichi Tatemura, and Belle L. Tseng. 2007. Structural and temporal analysis of the blogosphere through community factorization. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 163–172, New York, NY, USA. ACM.

K. Church and P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16:22–29.

J. Coleman, H. Menzel, and E. Katz. 1966. *Medical Innovations: A Diffusion Study*. Bobbs Merrill.

P. Domingos and M. Richardson. 2001. Mining the network value of customers. In *Seventh International Conference on Knowledge Discovery and Data Mining*.

E. Garfield. 1955. Citation indexes for science; a new dimension in documentation through association of ideas. *Science*, 122(3159):108.

E. Garfield. 1972. Citation analysis as a tool in journal evaluation. *Science*, 178(4060):471–479.

P. Gloor, J.S. Krauss, S. Nann, K. Fischbach, D. Schoder, and B. Switzerland. 2008. Web science 2.0: Identifying trends through semantic social network analysis.

G. Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA, USA.

D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. 2004. Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*, pages 491–501. ACM New York, NY, USA.

D. Hindle. 1990. Noun classification from predicate-argument structures. In *28th Annual Meeting of the Association for Computational Linguistics*, pages 268–275, Pittsburgh, PA. Association for Computational Linguistics.

T. Holloway, M. Božicevic, and K. Börner. 2007. Analyzing and visualizing the semantic coverage of wikipedia and its authors. *Complexity, Special issue on Understanding Complex Systems*, 12(3):30–40.

A. Kilgarriff, P. Rychly, P. Smrz, and D. Tugwell. 2004. The Sketch Engine. *Proceedings of Euralex, Lorient, France*, pages 105–116.

M. Kimura, K. Saito, and R. Nakano. 2007. Extracting influential nodes for information diffusion on a social network. In *Proceedings of AAAI*, volume 22, page 1371. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

J.M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632.

J. Kleinberg. 2003. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397.

R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. 2005. On the bursty evolution of blogspace. *World Wide Web*, 8(2):159–178.

J. Leskovec, L. Backstrom, and J. Kleinberg. 2009. Memetracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506. ACM New York, NY, USA.

D. Lin. 1998. Automatic retrieval and clustering of similar words. *COLING-ACL, Montreal, Canada*.

V. Mahajan, E. Muller, and F. Bass. 1990. New product diffusion models in marketing: A review and directions for research. *Journal of Marketing*, 54(1):1–26.

S. Morris. 2000. Contagion. *Review of Economic Studies*, 67:57–78.

S. Nakajima, J. Tatemura, Y. Hino, Y. Hara, and K. Tanaka. 2005. Discovering important bloggers based on analyzing blog threads. In *Workshop on the Weblogging Ecosystem, 14th International World Wide Web Conference*.

L. Page, S. Brin, R. Motwani, and T. Winograd. 1998. The pagerank citation ranking: Bringing order to the web.

P. Pantel and D. Lin. 2002. Discovering word senses from text. In *Proceedings of ACM SIGKDD02*.

F. Pereira, N. Tishby, and L. Lee. 1993. Distributional clustering of english words. In *Meeting of the Association for Computational Linguistics*, pages 183–190.

D. Rice. 1965. Networks of scientific papers. *Science*, 149:510–515.

E.M. Rogers. 1979. Network analysis of the diffusion of innovations. *Perspectives on social network research*, pages 137–164.

A. Rumshisky and V. A. Grinberg. 2009. Using semantics of the arguments for predicate sense induction. In *Proceedings of 5th International Conference on Generative Approaches to the Lexicon (Gl2009)*, Pisa, Italy.

H. Schutze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

D. Widdows and B. Dorow. 2002. A graph model for unsupervised lexical acquisition. In *19th International Conference on Computational Linguistics*, pages 1093–1099, Taipei, Taiwan.

H. P. Young. 1998. *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton.

T. Zesch and I. Gurevych. 2007. Analysis of the Wikipedia category graph for NLP applications. In *Proc of NAACL-HLT 2007 Workshop: TextGraphs*, volume 2.

V. Zlatić, M. Božičević, H. Štefančić, and M. Domazet. 2006. Wikipedias: Collaborative web-based encyclopedias as complex networks. *Physical Review E*, 74(1).