iZENEsoft Topic Graph Engine

iZENEsoft R&D Team

February 18, 2011

Contents

1 Requirements

iZENEsoft Topic Graph Engine is a package of data mining algorithms that dedicated on topic detection, topic tracking, topic link detection, and topic sentimental analysis. This package is summarized and encapsulated from a series of requirements, including:

1. Auto document hyperlinking
   SF1 will automatically insert hyperlinks directly into the document itself. This can be done to any granularity, such as a word, a phrase, or a sentence. By automatically identifying sub-areas of the document for which there are relevant links, SF1 creates a seamless delivery of related information to the user. As users read through documents, they can delve into or learn more about a concept by easily clicking on an automatically-generated hyperlink. Key benefits:
   - Reduces the cost of maintaining unstructured information
   - Reduces the time it takes to navigate to related information
   - Reduces duplication of effort
   - Keeps people informed and up-to-date
   - Retains browsers or consumers on a website by dynamically recommending related content and products
2. Auto query linking
   SF1 will automatically return topic hints that relevant to queries. This is different with two another utilities:
   - Taxonomy generation—which is used to filter query results using hierarchical taxonomies, or labels(which have some relatinship with topics within corpus).
   - Query recommendation—which is used to return relevant queries from query logs.
3. Buzz insight
   General BI refers to Business Intelligence, while according to our teminology, we use a trademark Buzz Insight instead. From the long run viewpoint, our BI product will be an eventual Social Media Intelligence solution, while at current stage, we define BI as a solution that analyzes online conversations and interactions for specific information—so called “buzz”. This product is used for corporations, brands, products, persons and top news occurring in the Internet communities. With powerful data analysis and mining capabilities, it can help you to make informed decision based on the refined reports shown in diverse formats.
4. Public opinion monitoring
   This project requires the burst topics could be discovered real timedly, and show polarity reports to administrators.

The above four requirements have consensus internal essentials:

- They all require to detect topics from corpus. The word topic has many other aliases when applied to different vertical areas, such as Buzz[6], Meme[9], Event[4], Incident[5], Bursts[7], or Key Phrases.
- They all require to detect the latent links among topics. Such a process may have some variations when applied to different areas, such as Event Threading[5], Topic Chaining[8], Topic Tracking[11], and Meme Tracking[9].

As a result, it’s reasonable to consider them together to provide high reusage. We name such an encapsulation as Topic Graph Engine, while after the Opinion Mining module is integrated in future, together with our search suites as well as Recommender Engine, it will form our `Social Media Business Intelligence` solution, which is the key product model in the quadrant that we gather public data and provide services to enterprise users(private). `Autonomy` has made intensive research and development on such quadrant, and they figure out these core components which will be our clear guidance on the product roadmap evolution.

- Social media connectors,which connect to a series of social networks including RSS, Twitter, Facebook, Yelp, LinkedIn, TripAdvisor, Yahoo! Finance, CNet Reviews, WebMD, IMDb, Kbb.com, and Epicurious.
- Sentiment analysis, which offers advanced categorization based upon degrees of sentiment and tonality. By analyzing the structures and meaning of language, it could determine the positive and negative characteristics of each piece of information and creates relevant classification systems. Marketers can apply multiple tagging functions and specific threshold cut-offs to determine the sensitivity of sentiment analysis. As the content is analyzed for sentiment, it is classified dynamically and in real-time, resulting in immediate benefits. Businesses can respond quickly to negative sentiment of their product as the opinions arise in the user generated content, or adjust their marketing campaign to affect positive reception. Moreover, Autonomy enables sentiment analysis of audio and video so that marketers can take advantage of the rich repository of information residing in multimedia social media assets. For instance, during playback, each speaker is marked in the media player with a different color, and the areas that contain heightened emotion are automatically marked during conversation. The combination of speaker separation, cross-talk identification, and emotion detection allows organizations to quickly identify and understand specific customer attitudes.
- Reputation analysis, which operates consistently on a large pool of changing information over a period of time to identify significant positive and negative trends in opinion. Organizations can measure how they are viewed by the ever-active social networks. This intelligence can then form the basis of an official response, ensuring the enterprise does not ignore this increasingly powerful soapbox. Reputation analysis automatically performs detailed statistical analysis to identify emerging trends and their implications for the reputations of people, companies, and products. It aids strategic communications programs across the spectrum of traditional and new media, including newspapers, television, blogs, forums, message boards, social networks, and online communities.
- Clustering. Autonomy groups disparate pieces of feedback automatically according to conceptual and/or sentiment similarity. In this way, it offers visibility over the consensus of authorial opinion, as well as onto individual communications. Marketers can easily react based on the prevailing opinions and trends that are forming. Users can also be grouped so that the marketer can learn which group is enjoying a compelling experience, and what group is disgruntled with the product.
- Visualization. Visualization of information can pre-empt the need to build complex models to identify trends. By clustering and displaying information, trends can be identified without months of building complex models or cubes. Autonomy provides numerous visualization tools to complement its content analytics and improve pattern detection. For instance, the Autonomy Spectrograph allows the marketer to follow consumer interactions from start to finish, as it displays sentiment changing and evolving over time. It also shows different points where conversations branch-off onto related subjects. In addition, a full range of statistical data is produced to help marketers make strategic decisions on the management of a brand or product.
- Audio & video analysis. Text analysis functions are also supported within video and audio files including concept extraction, clustering, hyperlinking, and sentiment analysis, guaranteeing marketers full access to all sources of social media information.

- Automatic hyperlinking, which is mentioned previously.
- Profiling. Social media analysis can accurately understand individuals' implicit interests based on browsing, content consumption, or content contribution. Generating a multi-faceted conceptual profile of each user based on concepts and pattern analysis of click-through and submission means they represent a very current understanding of the users' interests, with no need for explicit input of any form from the user. Profiling can take place across multiple devices. A profile generated through a user's interaction with PDA content can then recommend webcontent or news content via email or SMS.
- Monitoring & alerting. Autonomy produces alert notifications whenever a notable event occurs, whether important new content is posted online, a growing number of the same product praise or complaint is detected, or inaccurate information is disseminated to a large audience. The growth of social media may have enhanced corporate transparency, but it can also expose organizations to serious legal issues. Understanding of social media contents can be used to monitor the social networks and circumvent legal issues by alerting users to content that discloses sensitive material.
- Social Search.

From Autonomy's works we summarizes such an equation :
$SocialMediaBusinessIntelligence = SearchEngine + TopicGraphEngine + RecommenderEngine$

While each of them could be used separately both as a product and a `SaaS` cloud service.

2 Problem Definition

2.1 Topic Detection

Topic detection refers to the process of discovering topics from corpus automatically. Both `Topic Models` and `KPE` could be used for that purpose. Additionally, there are some variations for topic detections within products:

- Topic detection for auto document hyperlinking and auto query linking is a kind of `KPE` process.
- Topic detection for other applications requires a denoising process based on the outputs of `KPE` module.
  1. For BI, the topics are required to have more relationship with nouns as well as their attributes.
  2. Topic detection for public opinion requires topics to be real `opinions`. It contains `When`,`What`,`How` as well as a simple event description.
- Topic detection for public opinion monitoring focus more on `burst` topics, which are those occur within latest time window. It requires the detection module to be able to output burst topics from a text streams real-timely. Those topics that have a long time distance from current time will be ignored.

As a result, the `KPE` module will be more flexible and configurable to be suitable for more application context.

2.2 Link Detection

Link detection refers to the process of linking topics into a graph according to some certain similarity measure or relevance measure. Previously, we have TG module which is also a kind of link detection process that generates the hierarchical tree according to syntatic relatedness. The link detection in `Topic Graph Engine` has a more general definition on the graph as well as relevance measure, see in fig 1 and fig 2.

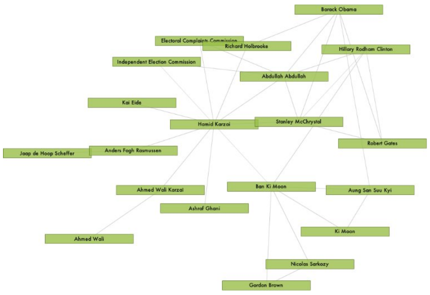


Figure 1: Topic Graph

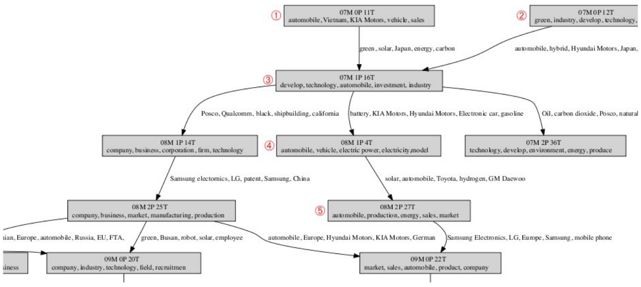

①
07M 0P 11T
automobile, Vietnam, KIA Motors, vehicle, sales
②
07M 0P 12T
green, industry, develop, technology,
green, solar, Japan, energy, carbon
automobile, hybrid, Hyundai Motors, Japan,
③
07M 1P 16T
develop, technology, automobile, investment, industry
Posco, Qualcomm, black, shipbuilding, california
battery, KIA Motors, Hyundai Motors, Electronic car, gasoline
Oil, carbon dioxide, Posco, natural
08M 1P 14T
company, business, corporation, firm, technology
④
08M 1P 4T
automobile, vehicle, electric power, electricity,model
07M 2P 36T
technology, develop, environment, energy, produce
Samsung electomics, LG, patent, Samsung, China
solar, automobile, Toyota, hydrogen, GM Daewoo
08M 2P 25T
company, business, market, manufacturing, production
⑤
08M 2P 27T
automobile, production, energy, sales, market
ian, Europe, automobile, Russia, EU, FTA,
green, Busan, robot, solar, employee
automobile, Europe, Hyundai Motors, KIA Motors, German
Samsung Electronics, LG, Europe, Samsung, mobile phone
iness
09M 0P 20T
company, industry, technology, field, recruitmen
09M 0P 22T
market, sales, automobile, product, company

Figure 2: Topic Graph

As a result, `TG` could be looked on as a module for `Topic Graph Engine`, while we need more variations for different application context:

- The relevance measure of auto document hyperlinking could directly borrow from `TG`, and we need adjustment to link those topics into a `DAG` instead of a tree.
- The relevance measure of auto query linking could also choose that of `TG`, while under this application context, we need to consider another dimension factor when constructing the `DAG`—temporal.

  We need to figure out two kinds of relevant topics for given queries:

  - Global temporal, which means the relevant topics within the topic graph does not consider the temporal information.
  - Local temporal, which means we need to compute topic graph within a series of time window. User could specify any time window during queries, while we return relevant topics from time windows that are included by user specified time range.

  Temporal-specific topic graphing is not a must for auto query linking, however it's a must for BI.

- As to BI, besides the temporal dimension mentioned above, we need another dimension for the topic graph construction—site dimension.
  BI is a service that need to crawl data from multiple social medias, which could be either shared by all customers, or specified by some certain users who hope to be private ones. So the topic graph engine should be able to provide:
  - Global sites topic link detection, which means the relevant topics within the topic graph does not consider the private social media information.
  - Local site detection, which means the relevant topics will consider both shared social media information and private information.
- The link detection requirements for public opinion is similar to auto query linking—only temporal will be considered during constructing the topic graph.

3 Extra Design Issues

3.1 Auto Document HyperLinking

Auto document hyperlinking has an extra design issue—present user friendly documents that have contained the inserted hyperlinks. The hyperlinks could be only inserted automatically according to the content from `SCD` documents, which have neglected some special characters, such as html tags. It's better to unescape those characters for more friendly presentation, additionally, snippet algorithms are required if `SCD` documents are too long.

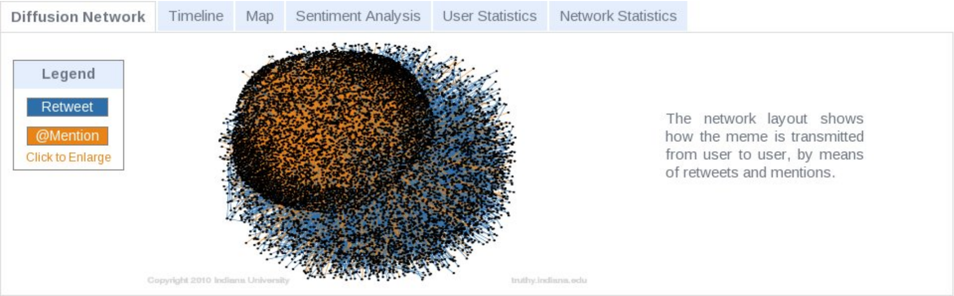
3.2 BI

According to definitions till now, we have made these changes to BI compared with existing solutions:

- Previously, `Buzz` is specified totally by users, and string matching algorithms are performed just after web pages are crawled. So BI will provide frequency statistics of those `Buzz` words within user specified time range. While with new BI, `Buzz` are automatically discovered from crawled data, besides the frequency statistics, we add another important intelligence dimension —relevant `Buzz`. Additionally, we also require the users to specify the `Buzz` words he most care, and the BI will find the most relevant `Buzz` chain for the given words. So this product is very close to eventual social media intelligence in that the opinion mining module would be included in future. From the implementation point of view, the BI system is very like existing `SF1` except that it contain some extra plugins—it means we implement new BI using `C++` instead of previous `Java`, while after `Recommender Engine` is finished, our `SF1` is also a server that provides full-stack solutions beyond search, just like `Autonomy IDOL` server does.
- Another important design issue for BI is data visualization. How to present data to users with a clear visualization is extremely important for the success of BI service. Either `Flex` or `Javascript` could be chosen for data visualization while it's far from enough—it requires a fully understanding over the data structure to be presented at UI layer, as well as a good imagination. Here are the information dimension that we need to make visualized:
  - Buzz frequency within any time window.
  - Buzz chain within any time window together with frequency statistics.
  - Buzz polarity(This dimension will not be provided until opinion mining is available).

Following are some data visualization effects for text stream mining:

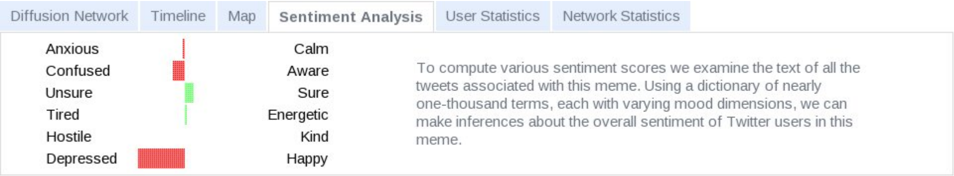
- Topic propagation in social media



The network layout shows how the meme is transmitted from user to user, by means of retweets and mentions.

Figure 3: Data Visualization of Topic Propagation

- Topic polarity(Notice that it has contained 5 categorizes, while our existing sentimental analysis can only provide 2—positive and negative)



To compute various sentiment scores we examine the text of all the tweets associated with this meme. Using a dictionary of nearly one-thousand terms, each with varying mood dimensions, we can make inferences about the overall sentiment of Twitter users in this meme.

Figure 4: Data Visualization of Topic Opinions

- Document browser based on topics[2].
- Event extraction with temporal references[10].
- Meme tracker[1].

All of the these examples are not well enough to be suited to our BI visualization, however, they could bring some hints.

- Implementation. Product team will rewrite BI UI based on current solution from Mar using Java. As we described above, there are some differences for new BI. These modules will be reused for UI:

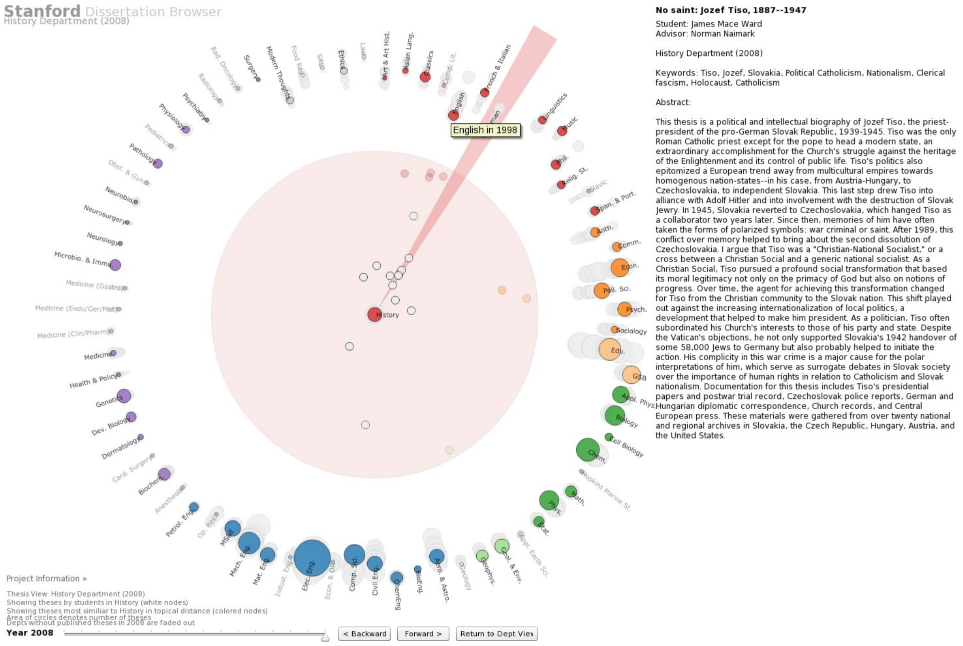
  - User management.
  - Site management.
  - User buzz words management.

  While these functions will not be used any more:

  - Data management—Old BI use data base to store all crawled data and buzz, while with new BI, all comes from BI server—with a similar architecture with SF1.

  And these modules are newly added for new BI:

  - Data visualization.
  - Buzz trend management.



Stanford Dissertation Browser
History Department (2008)

No saint: Jozef Tiso, 1887--1947

Student: James Mace Ward
Advisor: Norman Naimark

History Department (2008)

Keywords: Tiso, Jozef, Slovakia, Political Catholicism, Nationalism, Clerical fascism, Holocaust, Catholicism

Abstract:

This thesis is a political and intellectual biography of Jozef Tiso, the priest-president of the pro-German Slovak Republic, 1939-1945. Tiso was the only Roman Catholic priest except for the pope to head a modern state, an extraordinary accomplishment for the Church's struggle against the heritage of the Enlightenment and its control of public life. Tiso's politics also epitomized a European trend away from multicultural empires towards homogenous nation-states--in his case, from Austria-Hungary, to Czechoslovakia, to independent Slovakia. This last step drew Tiso into alliance with Adolf Hitler and into involvement with the destruction of Slovak Jewry. In 1945, Slovakia reverted to Czechoslovakia, which hanged Tiso as a collaborator two years later. Since then, memories of him have often taken the forms of polarized symbols: war criminal or saint. After 1989, this conflict over memory helped to bring about the second dissolution of Czechoslovakia. I argue that Tiso was a "Christian-National Socialist," or a cross between a Christian Social and a generic national socialist. As a Christian Social, Tiso pursued a profound social transformation that based its moral legitimacy not only on the primacy of God but also on notions of progress. Over time, the agent for achieving this transformation changed for Tiso from the Christian community to the Slovak nation. This shift played out against the increasing internationalization of local politics, a development that helped to make him president. As a politician, Tiso often subordinated his Church's interests to those of his party and state. Despite the Vatican's objections, he not only supported Slovakia's 1942 handover of some 58,000 Jews to Germany but also probably helped to initiate the action. His complicity in this war crime is a major cause for the polar interpretations of him, which serve as surrogate debates in Slovak society over the importance of human rights in relation to Catholicism and Slovak nationalism. Documentation for this thesis includes Tiso's presidential papers and postwar trial record, Czechoslovak police reports, German and Hungarian diplomatic correspondence, Church records, and Central European press. These materials were gathered from over twenty national and regional archives in Slovakia, the Czech Republic, Hungary, Austria, and the United States.

Project Information »

Thesis View: History Department (2008)
Showing theses by students in History (white nodes)
Showing theses most similiar to History in topical distance (colored nodes)
Area of circles denotes number of theses
Depts without published theses in 2008 are faded out

Figure 5: Data Visualization of Document Browser Based on Topics

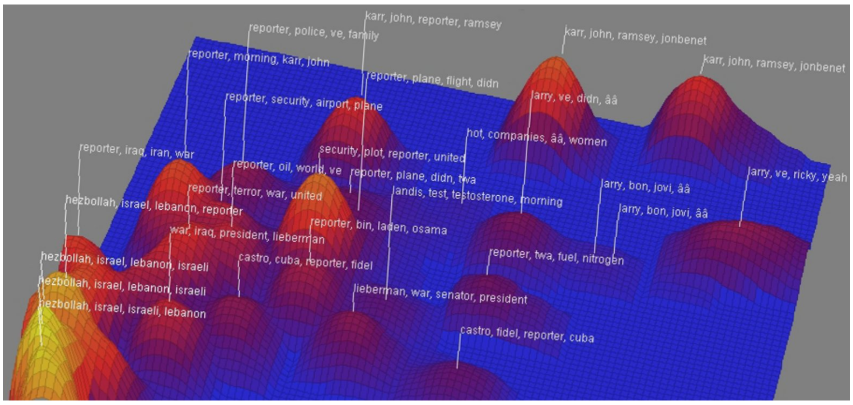


Figure 6: Data Visualization of Event Extraction With Temporal References

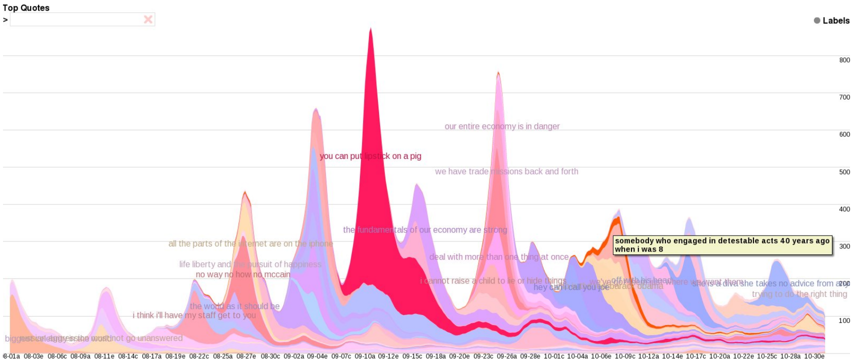


Figure 7: Data Visualization of Meme Tracker

3.3 Public Opinion

When BI is ready, public opinion could be ready soon in that we should guarantee these utilities to be ready:

- Bursts detection.
- Opinion mining.

Eventual public opinion solution will also perform as a SF1-like server with open-API exposed.

4 Future Application

`Topic Graph Engine` will be one of the three core engines within future `SF1` server. Besides the above mentioned four requirements, as well as the eventual social media intelligence solution, which is a core part of enterprise 2.0, it could bring extra benefits:

- Target trending Ads. Trending Ads are ads that are relevant in realtime to trending topics as they emerge. The ad units automatically update in realtime to show advertiser content that is relevant to trending topics as they emerge on social media. Oneriot[3] is a successful story for such application.

4.1 Schedule

Table 1: Topic Graph Engine

Milestone	**Due Date**	**In Charge**	**Description**
Auto Hyperlinking	2011-03-31	Jia, Yingfeng	Finish auto hyperlinking utility, open api, and first version of topic trend engine
Auto Hyperlinking Presentation	2011-04-15	Xin	Document browsing utility for auto hyperlinking
Auto Query Linking	2011-04-30	Jia	Finish auto query linking utility, open api
Topic Trend Engine	2011-05-30	Jia,Yingfeng,Jun	Topic trend engine without opinion is ready.
Opinion Classification	2011-06-30	Ben	Implement opinion classification module for BI and public opinion
BI	2011-07-31	Jia,Ben,Guangfeng	Finish new BI
Public Opinion	2011-07-31	Jia,Ben,Guangfeng	Finish public opinion

References

[1] http://memetracker.org/.

[2] http://nlp.stanford.edu/projects/dissertations/browser.html.

[3] http://www.oneriot.com.

[4] A. Das Sarma, A. Jain, and C. Yu. Dynamic relationship and event discovery. In Proceedings of the fourth ACM international conference on Web search and data mining, pages 207–216. ACM, 2011.

[5] A. Feng. Incident threading in news. PhD thesis, University of Massachusetts Amherst, 2008.

[6] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo. Deriving marketing intelligence from online discussion. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pages 419–428. ACM, 2005.

[7] D. He and D.S. Parker. Topic dynamics: an alternative model of bursts in streams of topics. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 443–452. ACM, 2010.

[8] D. Kim and A. Oh. Topic Chains for Understanding a News Corpus. Computational Linguistics and Intelligent Text Processing, pages 163–176, 2011.

[9] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking 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, 2009.

[10] D. Luo, J. Yang, M. Krstajic, W. Ribarsky, and D. Keim. EventRiver: Visually Exploring Text Collections With Temporal References. IEEE Transactions on Visualization and Computer Graphics, 2010.

[11] J. Makkonen. Semantic Classes in Topic Detection and Tracking. 2009.