Enterprise Discussion Analysis

Sara Rosenthal
Department of Computer Science
Columbia University
New York, NY 10027, USA
sara@cs.columbia.edu

Ashish Jagmohan
IBM Research
Yorktown Heights, NY 10598, USA
ashishja@us.ibm.com

ABSTRACT

Recent business studies have shown that social technologies can significantly improve productivity within enterprises by improving access to information, ideas, and collaborators. A manifestation of the growing adoption of enterprise social technologies is the increasing use of enterprise virtual discussions to engage customers and employees. In this paper we present an enterprise discussion analysis system which seeks to enable rapid interactive inference of insights from virtual online enterprise discussions. Rapid understanding is facilitated by extracting a hierarchy of key concepts, which represent a multi-faceted thematic categorization of discussion content, and by identifying high-quality thematic exemplar comments. The concept hierarchy and exemplar comments are presented through an intuitive web user-interface which allows an analyst to quickly navigate through the main concepts and the most relevant comments extracted from the discussion. We present a preliminary validation of system efficacy through user surveys provided to test users.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Subjects—*information filtering, selection process*

Keywords

discussion analysis; enterprise discussions; comment selection; concept hierarchy; social enterprise; web interfaces

1. INTRODUCTION

The use of social technologies for collaboration has the potential to significantly improve productivity within enterprises. An analysis by the Mckinsey Global Institute ([Chui et al. 2012]) has estimated that such technologies can yield a 20-25% improvement in productivity, by improving access to information and collaborators. This, in turn, translates into a US$1.6-1.9T benefit across several sectors within four major economies. Not surprisingly, an IDC study ([Thompson 2012]) reports rapid growth in the social enterprise software market, projecting a compound annual revenue growth of 42%, reaching US$4.5B in 2016.

A specific manifestation of the adoption of enterprise social technologies is the increasing use of enterprise virtual discussions to collaboratively harness knowledge, opinion, and innovation, from employees and clients. In this context, an often encountered issue is how to effectively cull insight from the large quantities of discussion material typically generated by such an event. A completely manual analysis process can be costly and even infeasible, depending on discussion size.

In this paper, we present a novel discussion analysis system which aims at enabling rapid interactive analysis of enterprise virtual discussions. Rapid extraction of discussion insights is facilitated through a multi-faceted thematic categorization of discussion content, and identification of high-quality thematic examples. To this end, the proposed system uses a variety of natural language processing techniques to decompose a discussion, or a group of related discussions, into an automatically extracted thematic concept hierarchy. Individual comments serve as atomic content entities, and are associated with one or more elements in the concept hierarchy. The concept hierarchy and categorized comment exemplars are presented through an intuitive web user-interface, which allows an analyst to quickly navigate through the main discussion themes and most relevant comments exemplars, and form a rapid understanding of the discussion content. We present preliminary validation of the efficacy of the proposed approach through a user survey provided to test-users for a real-world enterprise discussion.

2. RELATED WORK

There has been a significant amount of work in building concept hierarchies from text. Earlier work [Sanderson and Lawrie 2000, Sanderson and Croft 1999] has been performed on a group of documents across several topics, such as web pages. More recent work has explored concept hierarchies in discussions such as spoken conversations [Rashid et al. 2012], e-mail [Yang and Callan 2008] and social media [Selcuk et al. 2008]. Using a hierarchy as a visualization method for browsing and search has been used in several genres, such as music and image search. For example, a hierarchical menu on the results of a web search is used in [Sanderson and Lawrie 2000]. To our knowledge, we are the first to leverage concept hierarchy and related comment extraction for analyzing virtual enterprise discussions, and the first to build a web interface using a concept hierarchy as

CIKM'14, November 3–7, 2014, Shanghai, China.
Copyright 2014 ACM 978-1-4503-2598-1/14/11 ...$15.00.
http://dx.doi.org/10.1145/2578726.2578744 .

the general technique for visualizing an online discussion. The proposed technique, while designed for virtual enterprise discussions, has general applicability; it is well-suited to analyzing discussions such as those found in web forums.

The prior work which is perhaps most similar to our system is [Rashid et al. 2012]. This work provides an ontology of concepts, a word cloud of the discussion, and summaries. However, their approach analyzes meeting conversations as opposed to online discussions. There are also significant differences in the underlying technology. For instance, the concepts represent speakers and dialog acts as opposed to concepts extracted from text. It is also built on the sentence level as opposed to the post (or speaker) level which can cause a loss of information. Further they generate summaries rather than the proposed exemplar comment extraction. In terms of visualization, the previous work is built as a Java GUI application while our interface is a webpage built with HTML and JavaScript, which allows it to easily be integrated into online discussions and provide real-time analysis. We also provide several filtering options, including ones exclusive to online discussions, such as likes and votes.

3. DISCUSSION ANALYSIS SYSTEM

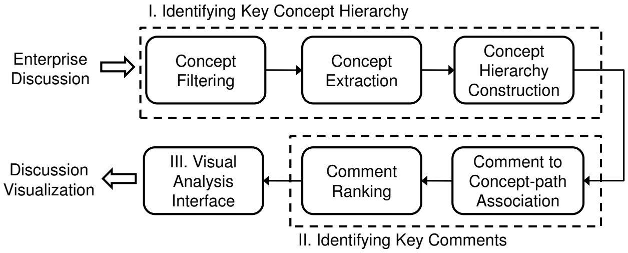

I. Identifying Key Concept Hierarchy
Enterprise Discussion
Concept Filtering
Concept Extraction
Concept Hierarchy Construction
Discussion Visualization
III. Visual Analysis Interface
Comment Ranking
Comment to Concept-path Association
II. Identifying Key Comments

Figure 1: Overview of discussion analysis system.

Figure 1 shows an overview of the discussion analysis system. The input to the system is the content of a virtual discussion. A typical discussion can consist of one or more topics or "questions" around which collective intelligence is sought. For example, a discussion question may be "How can your company increase use of analytics technology?". The discussion around a single question consists of a group of threads, where each thread is a set of several posts constituting a mutual dialogue. These threads can span over any time-range and may be synchronous or asynchronous, but are all related to the single question or topic. Our focus is on discussions conducted within an enterprise environment where clear questions have been asked of the participants.

The discussion content is supplied to the system through connectors for supported discussion platforms, or through flat-file content dumps. The system has three components. The first is the automated construction of a concept hierarchy through extraction of individual concepts, filtering to identify 'meaningful' concepts, and organization of identified concepts in a hierarchy. The second component is the identification and ranking of exemplar comments related to each concept-path in the hierarchy. Finally, the third component is the visual interface which presents the concept hierarchy and exemplar comments to the analyst, and allows interactive browsing and filtering. Each component is described in further detail in the following sections.

Table 1: Concept Filters

Filter	Description	Example
Stop Words	Exclude Lucene stopword list, custom stopwords	the, it
Plural Words	Merge plural and singular forms using WordNet	tools, businesses
Part-of-Speech	Include Noun, Noun-Noun, Noun-Adjective phrases	big data, analytics
HTML	Exclude html tags, urls	`www.web.com`
Gutenberg filtering	Exclude words in top 5% of Gutenberg corpus	use, need
Names	Exclude author names	Gary, Anne

3.1 Concept Hierarchy Identification

Concept hierarchy identification consists of three main steps (Figure 1): filtering the discussion text to identify potentially meaningful concepts, extracting a subset of significant concepts, and constructing a concept hierarchy from the extracted concepts. All information retrieval performed by our system is done via the Lucene package.[1]

A concept is a term that is considered to be meaningful in the discussion. Potential concepts are identified through the use of multiple filters, shown in Table 1. Terms or phrases of type noun, noun-noun and noun-adjective are considered as potentially meaningful concepts. Pluralization is detected using metadata from the WordNet [Miller 1995] lexical database, and singular and plural forms are merged. Stop-words, HTML tags and participant names are excluded. Also used is a 'Gutenberg' filter which excludes words which are very frequently found in the Gutenberg corpus;[2] very common words generally do not constitute meaningful concepts in business discussions, and this filter helps exclude such terms.

The filtered concepts are then ranked using a variety of metrics. Denote t as a concept term, d as a discussion from a set of discussions D, and q as a discussion-question from the set of questions Q in discussion d. Then, the following metrics are used to quantify the potential importance of the term t in discussion d:

- **Term Frequency** $tf(t,d)$ **and Term-Question Frequency** $tqf(t,q)$, the normalized number of occurrences of the term t in discussion d, and question q, respectively. This finds the most used terms in the current discussion and discussion topic.
- **Term Frequency-Inverse Document Frequency** $tfidf(t,d) = tf(t,d) \times idf(t,D)$ to find terms that are important to this discussion in contrast to all discussions.
- **Question Term Frequency-Inverse Question Frequency** $qtfiqf(t,q) = tqf(t,q) \times iqf(t,Q)$ to find terms that are important for this question in contrast to other questions within the discussion
- **Bigram frequency** $bf(t,d)$ which is zero if t is not a bigram, and is the normalized frequency of t in d otherwise. This specifically finds significant word pairs (e.g. "big data" or "machine learning"), which may be under-ranked by generic term-frequency metrics.

We gather the top m terms (typically $150 - 200$) for each metric and remove overlapping terms.

[1] `lucene.apache.org`

[2] `www.gutenberg.org`

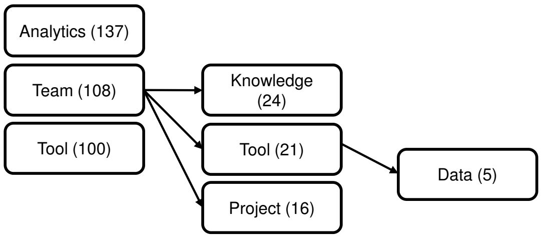


Figure 2: Example of hierarchical layout. Numbers in parenthesis indicate occurrence in query results of the current concept and its ancestor concepts (e.g. 'team + tool + data' occurs 5 times).

Finally, the gathered concept terms are organized into a hierarchy, for ease of exploration. Hierarchies have been found to be useful as a method for organizing information by the Library of Congress and the Dewey Decimal System. We automatically generate a hierarchy of concepts recursively in contrast to these manually created hierarchies. We initially order the filtered concepts based on the number of query results in Lucene. Each subsequent level of the hierarchy is constituted on the basis of the number of query results containing a given concept and all of its ancestors. In order to minimize the size of the hierarchical tree, we limit each level to show only the top n concepts (n can be configured by the analyst, typically, $n = 10$ or 15). Notably, we allow each concept to appear at several levels in the hierarchy. This is a parametric design choice (that can be disabled) which has been made to allow a user to explore complete alternate views of the concept hierarchy starting from any initial concept at the highest level of the hierarchy. Figure 2 shows an example hierarchy segment.

3.2 Identifying Exemplar Comments

The concept hierarchy represents a multi-faceted organization of the main thematic elements of the discussion. Next, we identify exemplar comments associated with each concept path, which, when presented to the end analyst, will allow the analyst to rapidly form an understanding of the discussion around that theme. We define a concept path as the list of concepts from the top level of the hierarchy to the current concept; a concept path thus identifies a fine-grained theme of discussion. An example is the concept path 'team -> tool -> data' shown in Figure 2. We find associated comments by querying the discussion for occurrence of the concept path terms in close proximity, using the SrndQuery class in Lucene. For example the query '(team) 10n (tool) 10n (data)' is used to find all comments with occurrences of the words team, tool, and data within a 10-word window (the window size is a configurable parameter), where the option 'n' indicates that the order of the words does not matter. Bigrams are queried using the 'w' option which constrains the words to be in order, e.g. '(big w data)'.

Ranking the comments for relevance at the most fine-grained level of the concept path is simple; we use the Lucene scores returned from running the SrndQuery for all concept path terms as described above. However, as we proceed up the concept path toward the root, the number of results increases and the SrndQuery scores become less meaningful from the perspective of relevance ranking. This is because the score is based off of number of occurrences, which may not translate into exemplar relevance. We hypothesize that comment relevance is related to how well the comment aligns with all fine-grained paths rooted at the current concept in the hierarchy. We quantitatively measure this heuristic for a comment by recursively summing the scores from each occurrence of the comment in lower levels of the hierarchy. Thus, for example, in Figure 3, comment 2 has the highest relevance score for concept C1, because it occurs most frequently on lower levels of the hierarchy tree rooted at C1. Ranking the comments in this manner ensures that comments which are well-aligned with multiple thematic facets associated with the current concept have higher rankings. The highest ranked comments are interactively displayed on the user interface.

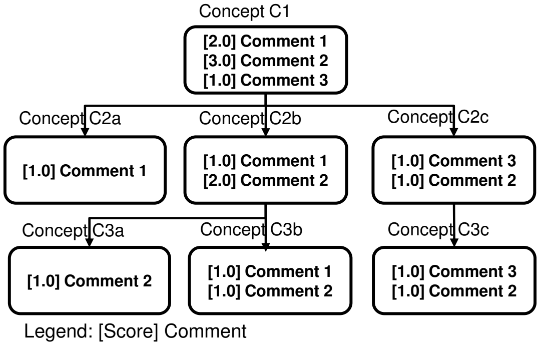
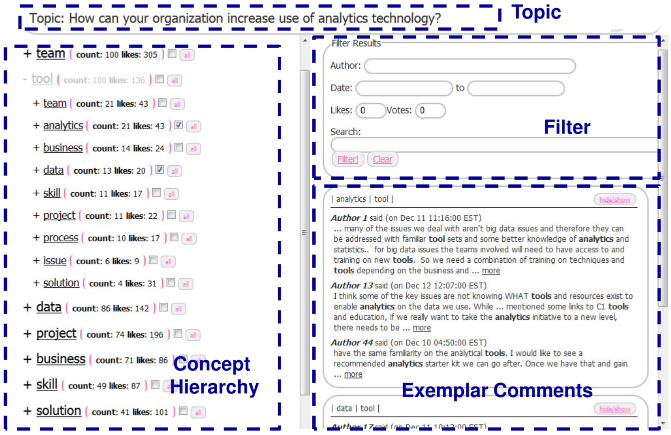


Figure 3: Combining comment scores for ranking.

4. USER INTERFACE



Topic: How can your organization increase use of analytics technology?

- + team (count: 100 likes: 305) □ all
- - tool (count: 100 likes: 136) □ all
  - + team (count: 21 likes: 43) □ all
  - + analytics (count: 21 likes: 43) ☑ all
  - + business (count: 14 likes: 24) □ all
  - + data (count: 13 likes: 20) ☑ all
  - + skill (count: 11 likes: 17) □ all
  - + project (count: 11 likes: 22) □ all
  - + process (count: 10 likes: 17) □ all
  - + issue (count: 6 likes: 9) □ all
  - + solution (count: 4 likes: 31) □ all
- + data (count: 86 likes: 142) □ all
- + project (count: 74 likes: 196) □ all
- + business (count: 71 likes: 86) □ all
- + skill (count: 49 likes: 87) □ all
- + solution (count: 41 likes: 101) □ all

Filter Results

Author:

Date: to

Likes: 0 Votes: 0

Search:

Filter! Clear

| analytics | tool | hide/show

Author 1 said (on Dec 11 11:16:00 EST)
... many of the issues we deal with aren't big data issues and therefore they can be addressed with familiar **tool** sets and some better knowledge of **analytics** and statistics.. for big data issues the teams involved will need to have access to and training on new **tools**. So we need a combination of training on techniques and **tools** depending on the business and ... more

Author 13 said (on Dec 12 12:07:00 EST)
I think some of the key issues are not knowing WHAT **tools** and resources exist to enable **analytics** on the data we use. While ... mentioned some links to C1 **tools** and education, if we really want to take the **analytics** initiative to a new level, there needs to be ... more

Author 44 said (on Dec 10 04:50:00 EST)
have the same familiarity on the analytical **tools.** I would like to see a recommended **analytics** starter kit we can go after. Once we have that and gain ... more

| data | tool | hide/show

Author 17 said (on Dec 11 10:12:00 EST)

Figure 4: User interface layout and example.

The concept hierarchy and extracted comments are presented to the analyst through a web interface built using HTML, CSS, and JavaScript. Figure 4 shows an example discussion analysis produced by the tool, overlayed with a description of the interface layout. The user interface consists of two main parts; the concept hierarchy (on the left), and the comment snippets (right). In addition, the topic associated with the discussion is shown at the top. The hierarchy is initially displayed as an ordered list of the top concepts. Each concept can be expanded to show the key concepts that are connected to it. Each level in the hierarchy can be expanded until all concepts have been reached.

Table 2: Top exemplar comments corresponding to example concept-paths for topic "How can your organization increase use of analytics technology?"

Concept Path	Top exemplar snippet
tool - analytics	"... many of the issues we deal with aren't big data issues and therefore they can be addressed with familiar tool sets and some better knowledge of analytics and statistics.. for big data issues the teams involved will need to have access to and training on new tools. So we need a combination of training on techniques and tools depending on the business ..."
skill	"While many of our analysts have stronger system and tool skills I think it would be beneficial to team them with [other] analysts who tend to have more business experience and knowledge ..."

When a concept on the left hand side is selected, the exemplar comments associated with it appear on the right hand side in the comments section. Several concepts can be selected at once by clicking on the checkbox to the left of each concept. There is also the ability to view the comments for all of the concepts. This enables the end analyst to rapidly gain understanding of the main discussion themes in breadth-wise and/or depth-wise fashion, as desired.

The comments can be filtered by several criteria; date, author, number of likes or votes, and key concepts. Comments and comment sets, corresponding to concepts, can also be sorted using one of multiple criteria. By default, the tool uses the novel information-based metric to rank comments described in section 3.2. The tool also allows sorting of comments by various meta-data metrics, such as the number of comment responses, comment likes, votes, etc. The content around each comment can be expanded to view the entire thread, which enables the user to quickly gain insight into the context of the comment.

The current version of the interface uses static html that is built from the discussion in advance. The advantage to this approach is that the user does not have to wait for most content to load. We use AJAX (notably JQuery) to do filtering on the fly. The web.py framework is used to provide certain server-side functionalities, for complex types of filtering such as customized sorting.

5. EVALUATION

In our experience with business discussions analyzed using the proposed tool, it generally does well at extracting important high-level concept-paths, and high-quality associated exemplars. Table 2 shows two examples[3] of the top exemplar found for two example concept paths for a real-world discussion analyzed by the tool. In each case the exemplars constitute high-quality suggestions which are holistically relevant to the discussion at hand. Thus, the first exemplar corresponding to the 'tool-analytics' path points out how training in analytics and tooling needs to be combined in the context of big data. The second exemplar, for the top-level 'skills' concept contrasts the team's system and tooling skills with business skills, and suggests a relevant course of action. The tool is less effective at identifying niche concepts which may occur infrequently, but may be important.

As a preliminary evaluation of the efficacy of the tool in aiding the user in understanding the discussion, we created a survey for analyst-users of our tool. The aims of the survey were to determine: 1. How useful is the tool in identifying important concepts and comments, and in providing an understanding of the discussion? 2. How easy is the tool to use for analysis? 3. How long did it take to use the tool to gain an understanding of the discussion? We asked detailed questions related to each of these aspects, on a Likert scale of 1-5. The questions dealt with the interface as a whole as well as specific aspects including understanding key concepts, finding comments, understanding the discussion, filtering the results, and tool design. The survey was filled out by 5 respondents who were each given the same discussion to analyze with the help of the tool. The majority of the respondents reported that they found the tool useful and easy to use; the overall usefulness of the tool and ease-of-use were each rated at the highest or second-highest level by four of the five users. In addition, the majority of respondents reported that the tool improved the speed at which they could understand the key concepts, find important comments, and understand the discussion.

6. CONCLUSION

We have described a discussion analysis system that can be used to gain insight into enterprise discussions. We have conducted surveys to determine its effectiveness, showing that users find the tool to be useful for rapidly gaining in-depth understanding of a discussion. The current interface is a strong foundation for understanding a conversation and there are many additions, such as opinion, that can be integrated to enhance user understanding of the conversation. We believe the system can also easily be adapted to include multiple discussions and other domains.

7. REFERENCES

[Chui et al. 2012] M. Chui and others. 2012. The social economy: Unlocking value ... *Mckinsey and Company, Insights and Publications* (2012).

[Miller 1995] G.A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38 (Nov. 1995), 39–41.

[Rashid et al. 2012] S. Rashid, G. Carenini, and R.T. Ng. 2012. *A Visual Interface for Analyzing Text Conversations.* Springer. 93 – 108 pages.

[Sanderson and Croft 1999] M. Sanderson and B. Croft. 1999. Deriving concept hierarchies from text. In *ACM SIGIR.* 206 – 213.

[Sanderson and Lawrie 2000] M. Sanderson and D. Lawrie. 2000. *Building, Testing, and Applying Concept Hierarchies.* 235 – 266 pages.

[Selcuk et al. 2008] K.C. Selcuk, L.D. Caro, and M.L. Sapino. 2008. Creating tag hierarchies for effective navigation in social media. In *ACM SSM.* 75 – 82.

[Thompson 2012] V. Thompson. 2012. IDC Worldwide Ent. Soc. Software 2012-2016 Forecast Upd. (2012).

[Yang and Callan 2008] H. Yang and J. Callan. 2008. Ontology Generation for Large Email Collections. In *Int. Conf. Dig. Govt. Res.* 254 – 261.

[3] Identifiable personal and business information has been removed from exemplars.