A Model of USENET Newsgroups Dynamics : Implementation and Results

M. Martinovic, G. Sampath, R. Wagner, S. Briening

Department of Computer Science
The College of New Jersey
2000 Pennington Road, Ewing, NJ, 08628, USA
{mmmartin, sampath, wagner8, brienin2}@tcnj.edu

Abstract: We present an implementation of the multilevel text processing model of discussions in USENET groups proposed earlier (NLDB '02 Proceedings). In the statistical processing phase, a discussion thread is SGML tagged to include the relevant information about parent-child relationships among the postings as well as other meta data of postings and threads. This tagged output is then processed by a generic information retrieval system. Various relevant metrics that measure properties of discussions (such as thread focus, relevance of posting, discussion density, etc.) are defined and computed. The subsequent semantic component (utilizing tools like electronic lexicons, POS taggers and parsers) has been implemented to work in a modular fashion to allow inclusion or exclusion of some of its subcomponents. The user may tune this module to its minimal level to process semantics of individual words only, or up to its maximal level to include words with their full contexts. We also present evaluation data assessing the performance of the system with or without some of its modules.

1. Introduction

In [SM02], a multilevel text processing model of discussions in newsgroups on USENET was presented. The primary objective of the model is to extract meaningful information from the discussions in a systematic fashion that utilizes the underlying structural and semantic properties of the discussion groups. Unlike in data mining, where structural information can be extracted from the data because such structure is an integral part of the stored data, text data generally has little associated structure ([WF99], [JTT99]). However, USENET discussions possess incidental structures in the form of metadata which is stored together with discussion threads. Such metadata considerably facilitates the task of information extraction, and the proposed model attempts to make extensive use of them. Mining for USENET is therefore based on some combination of: 1) extraction of keywords; 2) computing word associations; 3) identification of referents (names, places, data types); 3) soft parsing; and 4) statistics of word groups ([Ca97]). The work reported here is a continuation of the study begun in [SM02] and describes the detailed design and implementation of the model as well as an evaluation and results from its use on a sample set of discussions.

2. Design Techniques and Text Processing Tools

Our design approach is analogous to the IR system design approaches centered on corpus-based methods [NZ97]. Its current implementation consists of three top level modules and is depicted in Figure 1..

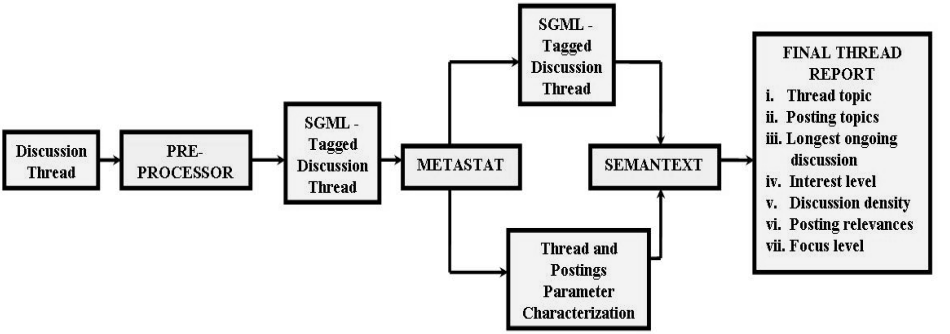

Figure 1: Implementation Diagram

The first module is a preprocessor that analyzes the original discussion thread file and inserts SGML tags into it. The tags identify and mark relevant meta data which are then conveniently exploited by the subsequent modules. The ensuing module METASTAT is extracting and further processing thread- or posting-related meta data recognized by the preprocessor. The values of pertinent thread and posting parameters are computed in this phase. The third module, SEMANTEXT, does a semantic analysis of the text contents in thread postings producing the characterization of the overall thread and posting topics.

3. Preprocessor Module

The preprocessing module downloads the messages from a news server via NNTP and then scans through the data identifying and marking the information that is newsgroup, thread and posting (message) related. The newsgroup-related tags include: **<NAME>**, **<DESCRIPTION>**, **<DATE_SPAN>**, **<ACTIVITY>**, and **<SOURCES>**. Related to each individual thread within the group, the module identifies **<THREAD_ID_LOCAL>**, **<SUBJECT>**,**<DATE_SPAN>**,**<ACTIVITY>**, and **<SOURCES>**. For each individual posting, **<SUBJECT>**,**<DATE>**,**<MESG_ID_LOCAL>**,**<MESG_ID_GLOBAL>**,**<PARENT_REFERENCES>**, **<CHILDREN_REFERENCES>**,**<THREAD_ID_LOCAL>**,**<THREAD_ID_GLOBAL>** are marked.

4. METASTAT Module

The next module called METASTAT extracts and computes the values of pertinent thread and posting parameters. Depth and breadth related metrics were retrieved from a generated thread structures. Once this information is obtained, the subsequent sections of

METASTAT use them to compute the values of parameters that characterize various features of the thread and its postings. We plan to implement a subject relativity measure based on the generic Information Retrieval System SMART ([Sa71]). Through various specification files, SMART is to be instructed which information is relevant to be extracted from the SGML tagged discussion thread files.

4.1 Feature Characterization Metrics

We introduce the following metrics collection in order to summarize and depict various relevant features of threads and their postings (messages) :

METRIC	FEATURE CHARACTERIZED
Maximal depth of the thread (md)	Most focused (longest ongoing) discussion of the thread.
Maximal breadth of the thread (mb)	Most provocative message (with most reactions).
Average depth of the thread (ad)	Focus level for the entire discussion thread.
Average breadth of the thread (ab)	Interest level in the topic over the entire thread.
Discussion density (dd) = (ad / md) * (ab / mb)	Discussion thread's density.
Posting irrelevance (pi) = pb / (pd * (pb+pd)),	Irrelevance of the posted message.
Posting relevance (pr) = 1 / pi	Relevance of the posted message.

Figure 2: Feature Characterization Metrics Table

4.2 Metrics Details

Maximal depth in a thread tree explicitly points out the longest ongoing discussion in the thread. In cases when participants are sincere about the exchange, this parameter can be used to indicate the level of the most focused discussion within the thread. Even when the above sincerity assumption is not met, one can say that the participants may have different foci and goals but the highest focus level cannot be denied. Maximal breadth of the thread reasonably well identifies the message that generated the maximal number of responses (the most *answer provocative* message). However, the most provocative message should not be confused with a message generating a rich follow up discussion. Discussion groups are abundant with outrageous postings that create a lot of immediate rebuttals but nothing more ('trolling', in newsgroup lingo).

The average depth of the thread is computed to include only the subthreads ending in a leaf node of the entire thread's tree. Their average depth characterizes how focused the entire thread is. The presence of very deep (very focused) subthreads improves the average while the presence and the number of very shallow ones decreases it. The average breadth of the thread is the mean value of the number of responses each message in the thread generates and is used to characterize the interest level. The interest level is proportional to the total number of responses to messages in the thread. The value of this parameter increases with an increased number of people responding but also with an increased number of responses from any individual participant. Since both identify interest level, the parameter seems fitting to depict the notion of interest.

Discussion density parameter reflects the thread's compactness and is proportional to normalized average depth and breadth of the thread. Thus, density of threads with few but very long subthreads (discussions) is appropriately characterized by a small value. Breadth-wise, threads with subthreads whose messages have a like number of replies would appropriately yield a larger value for the density parameter.

Relative irrelevance of a posting is devised to yield high values for postings with large breadth and small depth (a lot of immediate responses but few follow-ups). The same is true when both the breadth and the depth are small, which is indicative of messages with few immediate responses and few follow-ups. In other cases, the value of this parameter is appropriately small when describing messages with a lot of immediate responses and subsequent follow-ups, as well as messages with a few immediate replies but a large number of subsequent follow ups. Relative relevance is the reciprocal of the irrelevance and obviously not measured relative to the semantics of the corresponding message but to the environment's reaction to it. This fact is adequately addressed by the subsequent SEMANTEXT module.

5. SEMANTEXT Module

SEMANTEXT is the module that handles semantic analysis of the text contents of the postings. Its final product is a characterization for the overall thread topic, as well as for the subtopics in each of its messages. Its design facilitates experimentation with its submodules. Its modularity and portability enable moving, inclusion and exclusion of subcomponents while monitoring the performance of the system. The ultimate question that this approach is trying to address is how the use of NLP makes a difference in real-world applications that are similar. There are not very many studies which compare approaches that include NLP components with those that do not. One of the few that do is presented in [CNPB00]. Our evaluations are an attempt to provide more empirical data towards this goal.

Figure 3 depicts SEMANTEXT's modes of operation. Different color shades are used to represent known linguistic tools that our module uses in order to accomplish its goals (i.e. WordNet electronic lexicon, Eric Brill's part of speech tagger, ApplePie parser, end of sentence detector), as opposed to those that represent our "home grown" modules (i.e. MORPHNORM, CONTEXTOR, PSEUDONIM SELECTOR, STATCOMPUTER, STATCOLLECTOR, STATMERGER, SEMANSTAT). The WordNet tabs (SYN, HYPO, HYPER and COOR) suggest that the processing may or may not include usage of WordNet synonyms, hyponyms, hypernyms and coordinate terms. Similarly, the CONTEXTOR tabs indicate that the POS tagger, parser, and end of sentence detector can be used or omitted during the CONTEXTOR's execution.

The simplest mode of operation of SEMANTEXT (Process Designator (p.d.) α) is the one in which the text first goes directly into the morphological normalizer (MORPHNORM). There, the words get normalized by either stemming or lemmatization and are further passed into the next module (STATCOMPUTER).

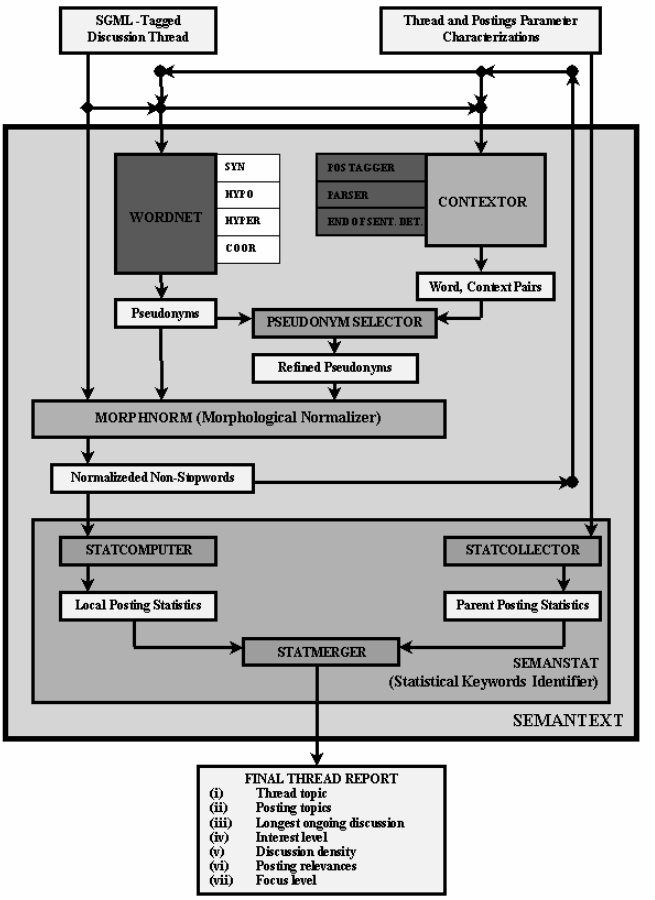

Figure 3: SEMANTEXT Modes of Operation

This submodule of SEMANSTAT computes statistics on words appearing in each message and passes them into the STATMERGER module. At the same time, another submodule of SEMANSTAT called STATCOLLECTOR collects statistics on phrases from the parent message(s) and provides them as additional input to the STATMERGER module**,** which combines them with the statistics of the currently processed message obtained from STATCOMPUTER. STATMERGER then characterizes the topic for each message by the top three key phrases in them. After the last message has been processed, the characterization of the entire thread is calculated. In addition, WordNet can be used (p.d. β) to produce pseudonyms (synonyms (p.d. β_1), hyponyms (p.d. β_2), hypernyms (p.d. β_3), and coordinate terms (p.d. β_4)) of the essential words in the current message. Those can then be passed into MORPHNORM directly or prior to that into the PSEUDONYM SELECTOR module. There they are combined with the context of the word to filter in the truly related pseudonyms and filter out unrelated ones. The set of refined pseudonyms is then passed into MORPHNORM module (p.d. γ). Furthermore, a

morphological normalizer can be used before WordNet and Contextor get to take on their tasks (p.d. δ). Of course, for WordNet this is only possible if MORPHNORM works as a lemmatizer and not as a stemmer.

On its part**,** CONTEXTOR can perform in three different modes itself, depending on which definition of the context it assumes. The simplest mode treats contexts as simple as ±n words surrounding the given phrase (p.d. γ_1). In the other two modules, the context is taken to mean the sentence in which the phrase appears. The sentence can be either used as a simple bag of words (p.d. γ_2) or as its parse tree (p.d. γ_3). The latter mode assumes the (rather computationally expensive) use of the parser.

The mode notation adopted here concatenates process designators to indicate that a particular corresponding submodule (process) was utilized during the processing. For example, $\alpha\beta\beta_1\gamma\gamma_2$ denotes the mode of operation in which the backbone process (α) was augmented to include the usage of WordNet for finding synonyms ($\beta\beta_1$), as well usage of CONTEXTOR with the context defined as the sentence's bag of words ($\gamma\gamma_2$).

6. Evaluation Procedure, Results and Outlook

Our first round of evaluations compared results obtained in different modes of operation of the system with the *correct* results obtained by human assessors. The comparison assesses the quality of produced characterizations, as well as execution times in different modes of operation. A correct characterization is awarded ***1/n*** points, where ***n*** denotes the ranking of the characterization. The data collection used in the experiment was obtained from *newscache0.freenet.de*, and is from the *alt.autos* discussion group, dated from 09/24/2002 to 10/24/2002. The tradeoff between the benefits of using NLP components is measured against the computational cost. Including WordNet improved the quality of characterizations on average by 1.4% (for processing descriptor β_1), 2.6% (β_2), 3.9% (β_3), and 3.9% (β_4), respectively. It also increased its execution time by 1.75 times (β_1), 2.75 times (β_2), 3.50 times (β_3), and 4.75 times (β_4) compared to the base mode, respectively. Incorporation of the CONTEXTOR module further improved the quality of the answers by an average of additional 0.7% (for p.d. γ_1), and 1.0% (for p.d. γ_2) but decreased it run-time performance by an average of 21.3 times (for p.d. γ_1), and 28.8 times when compared to the base mode (for p.d. γ_2), respectively. The table of Figure 4 summarizes the evaluation outcomes and seem to point out two major conclusions : (I) The incorporation of additional levels of NLP tools improves the accuracy of the system with an obvious and significant increase when coordinate terms and bag of words type context are included in processing. (II) The price for the increased accuracy of the system significantly reflects itself in a decrease of its computational performance. Notably, using coordinate terms and sentence context affected computation most seriously.

We see our present system as a framework for future augmentations justified by their evaluation results. Improvements are investigated for modules presently available based on what we learn during our experiments. In addition**,** the semantic analysis presently

does not go beyond the sentence level and is envisioned to incrementally include paragraphs, and hopefully entire messages and beyond. The issues of participant value and belief systems are on our future road map as well. We have yet to see if larger, more diverse and unexplored discussion groups might cause our system to behave differently.

Operation Mode	Success Rate	Run-Time Increase
BASE α	**+ 0 %**	*** 1**
SYN $\alpha\beta\beta_1$	**+ 1.4 %**	*** 1.75**
SYN HYPO $\alpha\beta\beta_2$	**+ 2.6 %**	*** 2.75**
SYN HYPO HYPE $\alpha\beta\beta_3$	**+ 3.9 %**	*** 3.50**
SYN HYPO HYPE COOR $\alpha\beta\beta_4$	**+ 5.3 %**	*** 4.75**
SYN CON +- $\alpha\beta\beta_1\gamma\gamma_1$	**+ 2.2 %**	*** 9.25**
SYN CON SEN+- $\alpha\beta\beta_1\gamma\gamma_2$	**+ 2.3 %**	*** 11.25**
SYN HYPO CON +- $\alpha\beta\beta_2\gamma\gamma_1$	**+ 2.9 %**	*** 23.50**
SYN HYPO CON SEN+- $\alpha\beta\beta_2\gamma\gamma_2$	**+ 3.5 %**	*** 3.50**
SYN HYPO HYPE CON +- $\alpha\beta\beta_3\gamma\gamma_1$	**+ 4.8 %**	*** 27.35**
SYN HYPO HYPE CON SEN+- $\alpha\beta\beta_3\gamma\gamma_2$	**+ 5.1 %**	*** 27.50**
SYN HYPO HYPE COO CON+- $\alpha\beta\beta_4\gamma\gamma_1$	**+ 5.7 %**	*** 28.95**
SYN HYPO HYPE COO CON SEN+-$\alpha\beta\beta_4\gamma\gamma_2$	**+ 6.2 %**	*** 29.25**

Figure 4: Evaluation Outcomes Table

7. Acknowledgements

This work is supported by NSF Grant EIA 0130798 and by the College of New Jersey.

Literature

[Ca97] Cardie, C. Empirical Methods in Info Extraction. In *AI Magazine*, 18:4, 65-79, 1997.

[CNP00] Cardie, C., Ng, V., Pierce, D., Buckley, C. Examining the Role of Statistical and Linguistic Knowledge Sources in a General-Knowledge Question Answering System. In *Sixth Applied Natural Language Processing Conference* (ANLP-2000), 2000.

[Fe98] Fellbaum, C., editor, *WordNet, An Electronic Lexical Database*. MIT Press, 1998.

[JTT99] Jiang, M-F., Tseng, S-S., Tsai, C-J. Discovering Structure from Document Databases. In N. Zhong and L. Zhou, editors, Methodologies for Knowledge Discovery and Data Mining. Lecture Notes in Artificial Intelligence 1574. Springer-Verlag, NY, 1999.

[NZ97] Ng, H.T., Zelle J. Corpus-Based Approaches to Semantic Interpretation in Natural Languge Processing. In *AI Magazine Winter 1987*, 18:4, 45-64 1997.

[Sa71] Salton, G., editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice Hall, 1971.

[SM02] Sampath, G., Martinovic, M. A Multilevel Text Processing Model of Newsgroup Dynamics. In *Proceedings of NLDB 2002 Conference*, Stockholm, 2002.

[SG97] Sekine, S., Grishman, R. Domain Project: Final Report. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*. 1997.

[St99] Staab, S. 1999. *Grading Knowledge: Extracting Degree Information from Texts. Lecture Notes in Artificial Intelligence 1744.* Springer-Verlag, New York, 1999.

[WF99] Witten, I. H., Frank, E. 1999. *Data Mining*. Morgan Kaufman, San Francisco, 1999.