Multi-Faceted Information Retrieval System for Large Scale Email Archives

Jukka Perkiö, Ville Tuulos, Wray Buntine
Helsinki Institute for Information Technology
P.O. Box 9800
FIN-02015 HUT, Finland
{firstname.lastname}@hiit.fi

Henry Tirri
Nokia Research Center
P.O.Box 407
FI-00045 NOKIA GROUP
henry.tirri@nokia.com

Abstract

We profile a system for search and analysis of large-scale email archives. The system builds around four facets: Content-based search engine, statistical topic model, automatically inferred social networks and time-series analysis. The facets correspond to the types of information available in email data.

The presented system allows chaining or combining the facets flexibly. Results of one facet may be used as input to another, yielding remarkable combinatorial power. In information retrieval point of view, the system provides support for exploration, approximate textual searches and data visualization. We present some experimental results based on a large real-world email corpus.

1 Introduction

We focus on a niche application of information retrieval and data mining, large-scale email archives. A remarkable amount of information resides in publically available archives of mailing lists. The archives are typically accessible through the Web via an interface that supports browsing emails and some rudimentary search operations. Considering the sheer amount of information in these archives and lack of link structure that helps ranking with ordinary web pages, the current interfaces can be considered suboptimal for harnessing all the buried knowledge. However, for a casual user the simple and familiar interface is often enough. In this paper we consider an approach which gives the user remarkably more powerful tools with only a moderate cost in usability.

The presented approach can be considered as a methodological testbed showing how various sophisticated information retrieval and data mining tools might work in concert. Emails are an attractive target for the system since they are multi-faceted by nature: They have a textual, temporal and a social dimension. All these facets are utilized by the presented system. Attractiveness of email data is also increased due to its abundance and practical importance of efficient email processing. Some prior work on sophisticated email management include [12, 7]. In contrast to many prior work which focus on personal mailboxes, we are interested in analysis and data mining of large-scale corpora.

All the facets presented in this paper have been developed and used by our research group previously, yet not in a multimodal context like this one. Our search engine techniques are published in [2, 13], the topic model based on multinomial principal component analysis (MPCA) e.g. in [3], topical trends in [8, 9] and social networks and evolution of relationships in [14]. These papers provide more technical details on the methods and empirical evaluations of performance.

Novelty of this paper lies in tight combination of these facets, making them a working and practical system. It appears that the seemingly unconnected methods form a powerful and seamless combination. No single facet could serve all different information retrieval needs adequately.

2 Facets

2.1 Search Engine

We have developed a full-fledged search engine which data structures are tailored for large-scale content-based ranking of textual data. The following two requirements were central in design of our ranking scheme: First, the user must be able to use inexact queries, meaning that no exact matches of query words are required although they are preferred. Secondly, we want to support *lazy queries* i.e. queries containing only a few words. We can not assume that the user is able to provide us a lengthy passage of text, representative of her needs.

The former desideratum is easily fulfilled with some probabilistic language model. We could use our topic model for ranking, as exemplified in [2]. However estimating topics based on only a few words, as required by our latter requirement, is somewhat unreliable. Moreover the topics do not always match with the user's needs, so there is need for a more fine-grained solution.

A standard method in information retrieval to tackle with lazy queries is automatic *query expansion* [1]. It possible to capture synonyms and other nearby concepts for a word by expanding it with the words which co-occur with it in the corpus. This method turns out to be too crude, losing specificity of the word, if applied as such. However when combined with proper normalization and especially with exact keyword search, it satisfies our first requirement. In addition, we segment each email to multiple constant-sized windows or blocks which are scored independently so that long and relevant emails are distinguished from long and irrelevant ones.

Let ϕ denote a set of attributes which are expanded with their co-occurring attributes. We call ϕ *ranking cue*. Each attribute w in the corpus, typically a word, is given a score as follows

$$S_w(\phi) = \frac{n(\phi, w)}{n(w)}$$

where $n(\phi, w)$ denotes the number of blocks in which ϕ and w co-occur and $n(w)$ denotes the number of blocks in which w occurs. Note that the formula is exactly conditional probability $P(\phi|w)$.

Let $\mathcal{B}_e$ denote the set of blocks in email e. Now each email is scored as follows

$$S_e(\phi) = \frac{1}{|\mathcal{B}_e|} \sum_{B \in \mathcal{B}_e} \sum_{w \in B} S_w(\phi)$$

thus it is the linear sum of scores of words in its blocks, normalized with number of blocks. Our content-based search works so that the user inputs a set of keywords and a cue set ϕ. For instance, in query "George bush /foreign /politics" keywords are "George bush" and ϕ ="foreign politics". The system retrieves all emails matching "George bush" using its inverted index, scores the resulting emails with $S_e(\phi)$ and returns the results ordered by descending scores.

We have tried to keep the search, and the system in general, as transparent as possible so that the user is able to maintain a mental model on *how* she could find a specific piece of information if needed, even though she does not have the full knowledge of the collection. For this purpose the search engine is the central tool due to its combinatorial power to produce subsets of the corpus. On the other hand the explorative and visualization facets allow us to keep the recall of emails high, since there is no need to make overly specific queries due to burden of browsing the results.

2.2 Topic Model

The topic model we use is based on a recent discrete or multinomial version of Principal Components Analysis (MPCA). These so-called multi-aspect topic models are statistical models for documents that allow multiple topics to co-exist in one document. They are directly analogous to the Gaussian basis of PCA which in its form of Latent Semantic Analysis (LSA) has been extensively explored in the text analysis community, but is not as well used in applications. Several kinds of experiments report MPCA methods have superior statistical properties to LSA, and the resultant components are also easier to interpret, see e.g. [6].

The simplest version of MPCA consists of a linear admixture of different multinomials, and can be thought of as a generative model for sampling words to make up a bag, for the Bag of Words representation for a document [1].

- We have a total count L of words to sample.
- We partition these words into K separate topics or components: $c_1, c_2, ...c_K$ where $\sum_{k=1,...,K} c_k = L$. This is done using a hidden proportion vector $\vec{m} = (m_1, m_2, ..., m_K)$. The intention is that, for instance, a sporting article may have 50 general vocabulary words, 40 words relevant to Germany, 50 relevant to football, and 30 relevant to people's opinions. Thus L=170 are in the document and the topic partition is (50,40,50,30).
- In each partition, we then sample words according to the multinomial for the topic, component or aspect. This is the base model for each component. This then yields a bag of word counts for the k-th partition, $\vec{w}_{k,\cdot} = (w_{k,1}, w_{k,2}, ..., w_{k,J})$. Here J is the dictionary size, the size of the basic multinomials on words. Thus the 50 football words are now sampled into actual dictionary entries, "forward", "kicked", "covered" etc.
- The partitions are then combined additively, hence the term admixture, to make a distinction with classical mixture models. This yields the final sample of words $\vec{r} = (r_1, r_2, ..., r_J)$ by totaling the corresponding counts in each partition, $r_j = \sum_{k=1,...,K} w_{k,j}$. Thus if an instance of "forward" is sampled twice, as a football word and a general vocabulary word, then we return the count of 2 and its actual topical assignments are lost, they are hidden data.

This is a full generative probability model for the bag of words in a document. The hidden or latent variables here are $\vec{m}$ and $\vec{w}$ for each document, whereas $\vec{c}$ is derived. The proportions $\vec{m}$ correspond to the components for a document, and the counts $\vec{w}$ are the original word counts broken out into word counts per component.

We have used Gibbs sampling [6] to learn these models from data using our own implementation of the model, available as an open source package[1].

2.3 Topical trends

For data that has inherent time structure, such as emails, it is appealing to apply MPCA model so that the temporal structure of the components can explored [8]. In particular we are interested in the temporal behavior of the latent variable $\vec{m}$.

As new emails arrive their topical representation i.e. distribution over the topic space is calculated, hence we get $\vec{m}$. The topical trend time series is estimated by creating a histogram for each component, each bin being a point in the time series. Bins are added depending on the desired resolution. The relative number of documents in component k at bin t is given by $\sum_i 1_{b_i=t} = m_{i,k}$, where b_i is the bin number for document i and $m_{i,k}$ is the proportion of document i in the component k. This is a real dynamic time series based on a fixed MPCA model. The model itself can be updated as needed.

Generally time series acquired this way are noisy and some sort of smoothing is needed. Well-known moving average methods are adequate for this purpose.

2.3.1 External model for topical trends

One interesting way of estimating topical trends is to use a pre-built model, against which new data is projected. This is beneficial if the user is more familiar with the pre-built model than the unseen data. We may also assume that the topics are named before, which further facilitates the trend analysis using topical trends.

In general the external model should be compatible with the new data. A model built from a corpus that mainly discusses flowers is not very useful for documents that are mainly related to something completely different, say information retrieval. The external model should either follow the same specific topic as the data to be projected or it should be very extensive and general in nature. The latter alternative suits well to our purposes. Our external models are based on Wikipedia[2], the free encyclopedia, that has comprehensive coverage on various subjects.

Estimation of topical trends over a corpus using an external model is not different from using a model based on the corpus itself – the most remarkable, and desirable, difference is that not all topics of the external model might be present in the corpus. For each document its topical distribution is calculated relative to the model and based on the desired temporal resolution a histogram is created, normalized and smoothed.

2.4 Social Networks

We are interested in social relationships that emerge from email communication. We call *social network* a graph in which nodes represent persons and edges relationships between them. There are several ways to automatically infer the edges. The most straightforward method is to induce an edge between two nodes (persons) if we observe an email sent from the first person to the second. We may naturally add a weight to each edge by observing the frequency of exchanged emails. At least in visualization point of view this approach has the downside of producing dense graphs with many weak, sporadic connections.

To make the edges to represent relationships more robustly, we may use email *threads*, namely long repeated sequences of emails on the same subject, as the edge generator. Thread is started by a message which does not relate to any previously active thread. A message is related to an active thread if its subject line equals to the subject of a known thread, excluding the possible 'RE' prefix, within some constant time-window. If two persons take part in the same thread, we induce an edge between them.

3 System

The facets are most efficient when used in concert. Table 1 shows modes of operation that emerge from pair-wise combinations of the facets. Each mode is described in detail in the following section. Even though the modes may appear rather complex, implementation builds upon a few basic elements. All search operations, the first row in the table, are based on our search engine. Each facet works as a special attribute which is used as a cue for ranking, as described in section 2.1. Every search operation produces a set of emails, possibly an empty set. Email set is the basic element upon which further analyzes, such as exploration, topical trends or further searches are based. Each facet accepts a set of emails as input and produces possibly a smaller, filtered, set of emails. Each operation may enrich emails with additional information which is typically used in visualization. For instance, estimated topic distribution of emails of a person may be used to shade the corresponding node in the social network.

The order in which the operations are applied affects the results. For instance, exploration including topic filtering followed by author-based search may produce different results than the same operations applied in reverse order. Note that since the operations may be easily chained, the combinatorial power of the system exceeds the simple pair-wise combinations presented in Table 1.

[1] http://cosco.hiit.fi/search/MPCA/

[2] http://www.wikipedia.org

	Search	Topics	Social Net	Time
Search	Content-based search	Topic-based search	Author-based search	Zeitgeist search
Topics		Exploration	Interest profiles	Topical trends
Social Net			Global relations	Evolution of relationships
Time				Browsing

Table 1. Facet combinations and the corresponding modes of operation

3.1 Modes of Operation

Content-based search allows the user to search textual contents of emails. If the emails of interest contain a known distinctive keyword, the desired emails may be found with simple keyword search. In case that no keyword characterizes the emails adequately or the keyword yields too many results, one may specify some *ranking cues* ϕ, namely some additional words, to the system. The cues do not affect the number of results returned but the results are returned in such order that the top ranking results contain maximum amount of cues or words which often co-occur with them in the corpus. For instance, with content-based search it is possible to find all emails talking about George W. Bush in the context of foreign politics – although no exact words "foreign politics" are mentioned in emails. Thus content-based search serves the information retrieval needs which are expressible with short textual queries.

If the user is not familiar with the corpus, forming effective content-based queries might prove difficult. In this case **topic-based search** functions as an automatic librarian by providing a familiar topic structure which can be used to guide searching. Content-based queries may return unexpected and confusing results if the user's own "mental model" of the corpus is inadequate. With topic-based search the user can easily grasp the prominent topics and use them to rank the results at broad level. One could focus searches, say, on physics. Topic-based search is bound to a predefined set of topics, since now cues ϕ are the topics, in contrast to the content-based search which allows more fine-grained, unconstrained queries.

Both content- and topic-based searches rely on textual contents of emails. Consider that the user is interested in all emails sent or received by a certain person. **Author-based search** solves these use cases trivially by providing a facility to retrieve a set of emails given the email address of sender or recipient. However we may further utilize the sender-recipient relationships: It is possible to use a person as the cue ϕ for content-based search. In effect, this lets the user to search for emails which are content-wise similar to the ones which are either sent or received by the specified person. This form of author-based search lets the user to use *shared interests* or *trust* as the basis for content mining. For instance, one might want to find emails about Linux having similar content and style as the emails written by Linus Torvalds – considering Torvalds' messages as authorative on the subject. Note that we may extend the effect by expanding ϕ to cover also the neighborhood of the specified person in social network. In a weighted graph, we may additionally drop weak links in the extended cover for increased robustness. For previous work on the subject, see e.g. [11, 5].

It is possible to retrieve all emails of some exact timespan using the aforementioned keyword search. Consider that a certain discussion or *thread* seems interesting to the user as whole and she would like to retrieve emails resembling the specified thread. Not unlike the previous author-based search it is also possible to use time as the cue ϕ for ranking. We call this function **Zeitgeist search** since it retrieves emails which resemble the ones written during a specified period of time. On public mailing lists, there are seldom more than a few active topics of discussion at the same time. Thus if one finds an interesting discussion spanning multiple emails and possibly several discussion threads as well, using Zeitgeist search it is possible to find discussions elsewhere having similar content.

MPCA topic model as such is best suited for **explorative** data mining. The emails are projected or classified under respective topics. The user may focus on topics in which she is specifically interested, or examine the global structure of the corpus by focusing on topic summaries. If we are using an external model, which is not based on the corpus itself, it is possible to find out which topics are *not* active in the corpus. For instance, one might notice with a glimpse that no email talks about biology in a specified mailing list. Finding this out with textual search would be prohibitively difficult.

We assume that all emails either sent or received by a person approximately represent her interests. By estimating active topics in this representative set of emails using the topic model, we may estimate the **interest profile** of the person. We may compare and match interest profiles of several persons. People sharing similar interests are easily noticeable from a social network in which nodes (persons) are shaded with a hue corresponding to their most prominent topic of interest.

Topical trends show change of topics in time. At broad level, we may examine e.g. dynamics of a full mailing list (see figure 2). More detailed analysis is possible by examining trends of a subset of emails, for instance the result set of a textual query, with respect to a larger set of emails. For instance, one might see how the usage of term "Al Qaeda" relates to topic "Middle East" during last few years. Since the graphs follow general word usage statistics in emails, giving them absolute interpretations is difficult. The intended usage is to see relative differences in general trends between various graphs. Topical trends may be also used to navigate to a specific discussion which would be burdensome to find otherwise amidst a long span of emails.

Social network spanned by emails show sender-recipient relationships. The network, or some part of it, may be visualized to show **global relations** between persons. The network helps to identify *hubs* in a set of emails i.e. persons having a large number of distinct connections. We may also visualize strength or activity of a connection (edge) between persons (nodes) or shade the nodes according to various attributes, such as interest profile mentioned above.

We may restrict the above network to a certain time-span. This allows us to show **evolution of relationships** by animating the network growth in time. Especially we may visualize how the connection between a pair of persons becomes active for a certain period of time and then fades away.

By **browsing** we refer to the familiar list of emails in chronological order, as typically used in personal email clients. This is the basic view for manipulating sets of emails.

4 Evaluation

Evaluating performance of a system that contains as many functional parts as the one presented in this paper is a complex task. As for multi-faceted systems, each facet can be evaluated individually, which alone can give some information about the system. Nevertheless the main idea of the presented system is seamless integration of each facet to effective information retrieval system. It is not sufficient to evaluate only the individual facets but combination of them working together. In the following we present some preliminary examples how the modes of operation work with respect to real-world data.

4.1 Data

We kindly got a snapshot of The Mail Archive[3] in November 2003. The dataset consists of some 20GB of public mailing lists of various subjects. Each email contains a time-stamp, the sender's email address, a subject line and the message body. Not all the mailing lists are in English, some contain vast amounts of automatically generated messages and some messages include binary attachments.

For the search engine functionality we indexed all the emails from which we could parse a proper time-stamp and a subject line. This resulted to a collection of 2,561,429 emails. We allowed very liberal case-insensitive tokenization which found 4,459,139 different tokens. Furthermore we tagged the time-stamp, thread-id, sender-id and thread-initiator-id and list-id for each email. The language of each email using TextCat [4] and all the sensible tokens appearing in emails written in English were lemmatized with the Porter stemmer [10]. The process took 2.5 hours in a dual processor 2.8GHz Xeon. The index contains tokens and stems in each email, their positions and the inverted index takes 3.3GB.

To guide the statistical topics produced by MPCA to semantically interpretable direction, we did some further preprocessing for the model. We first removed common stop words and applied part of speech tagging using TNT[4] part of speech tagger. After this we retained only nouns, verbs, adjectives and adverbs. This produced a lexicon of about 130,000 words with relatively good quality. The model was built using a multinomial for each of the retained word classes.

4.2 Topic Models

We used two different approaches for building the models. Firstly we built a 10 topic flat MPCA model for each mailing list. Secondly we used a pre-existing MPCA model that was built from the English Wikipedia corpus of June 2005 that contains some 600,000 documents. This model is a 100 topic flat model and all the topics are named manually.

The rationale of using two models is to see how the mailing lists relate to a well defined extensive model with pre-named topics. Another benefit is that the two models can be seen as a hierarchical way for analyzing the data. More general model (i.e. Wikipedia) is at the higher level in the hierarchy whereas the model built from the actual data provides more resolution and more exact description of the data.

4.3 Examples

A social network inferred from a mailing list is shown in Figure 1. The grey edges show the **global relationships** between the persons; they are inferred using the whole corpus. The network shows some interesting structures, especially a few highly-connected *hubs* in the center — these

[3] http://www.mail-archive.com/

[4] http://www.coli.uni-sb.de/∼thorsten/tnt/

correspond to exceptionally active participants of the mailing list. The black edges show *local relationships* among all the emails that match to a query. This is a particularly powerful feature to find out tacit correlations between email contents and persons.

For **explorative** use our system provides topics of the MPCA model. Table 2 shows the 15 most important lexemes for topics of a model that was built for the ctrl@listserv.aol.com mailing list. Solely by inspecting the topical keywords it may not be easy to differentiate between all ten topics. To better understand the model of this homogeneous corpus, one should also investigate the most prominent documents for each topic. Unfortunately it is not possible to show email messages due to lack of space. Because of the theme of the mailing list many topics are related to war, politics and social issues. Perhaps topics 3, 4 and 6 are the easiest to understand whereas topic 0 seems to be the least definite.

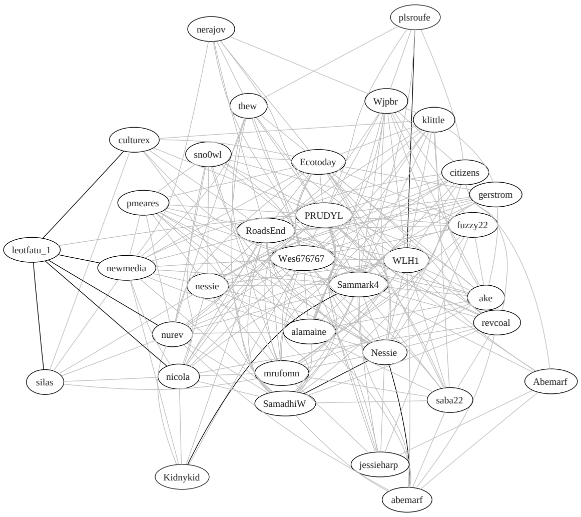

Figure 1. A social network showing global and local relationships

Examples of **topical trends** are shown in Figures 3–5. These trends are from the model that was built from the ctrl@listser.aol.com mailing list. Descriptions for the trends can be seen in Table 2. Figure 2 shows the temporal behavior of the 15 most prominent topics of the Wikipedia model in the same mailing list, rest of the topics were practically inactive in this mailing list. In the figure dark shades denote active topics. The corresponding names for the topics can be found in Table 3. These topics are very well in line with the topics that were estimated from the mailing list as can be seen when comparing Tables 2 and 3.

To experiment with the combination of **content- and author-based** search and topical trends we performed three kinds of queries to the mailing list data:

Topic ID	15 most frequent lexemes in relevance order
0	state; community; north; research; September; prison; discussion; agreement; police; group; company; search; report; fund; judge;
1	reply; matter; place; control; available; member; report; work; source; interest; book; end; people; war; direction;
2	theory; time; state; lector; government; research; archive; bush; endorsement; nazi; spectrum; post; holocaust; thought; caveat;
3	conspiracy; war; Clinton; u.s.; people; city; life; disclaimer; university; matter; action; group; us; Jew; center;
4	effect; new; president; system; officer; support; court; effort; reason; people; number; research; issue; business; john;
5	u.s.; group; agent; election; information; security; research; us; world; Israel; act; people; defense; position; house;
6	way; people; policy; war; attack; America; right; program; company; plan; available; Washington; course; misdirection; group;
7	order; west; company; end; matter; world; control; president; bank; video; light; business; security; part; committee;
8	world; official; york; people; credence; matter; CIA; u.s.; war: disclaimer; history; effect; group; rule; president;
9	people; time; available; information; official; disclaimer; intelligence; Israel; security; Iraq; letter; threat; world; movement; point;

Table 2. Fifteen most important lexemes for a ten topic model for the ctrl@listserv.aol.com mailing list.

1. General queries, which show topical dynamics of the query results in relation to the topic model.

2. An individual email address as a constraint to the search so that we can investigate how the behavior of one individual relates to the topical dynamics of the mailing list.

3. Using a subgraph of the social graph as constraint to the search so that we can investigate how the behavior of a “social group” relates to the topical dynamics of the mailing list.

Three example queries and their results’ temporal behavior are shown in figures 3 – 5. Trends that are

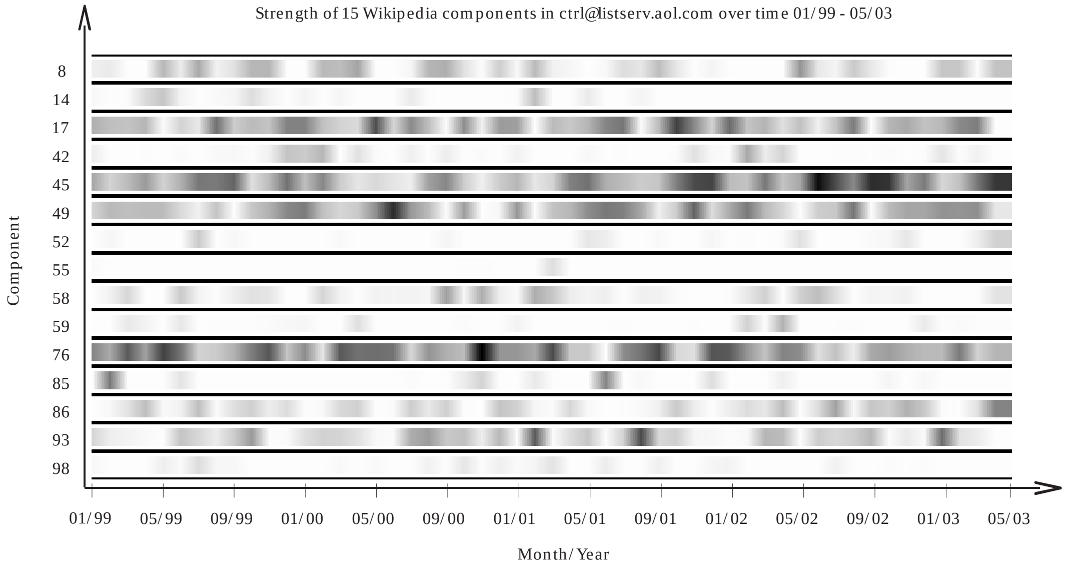

Figure 2. Temporal strength of 15 Wikipedia topics in the mailing list ctrl@listser.aol.com between 01/99 – 05/03.

Topic ID	Topic name
8	Economy
14	Physics
17	Computer Systems
42	Religion
45	Middle East
49	Theories
52	Military
55	Rock Music
58	U.S. Government
59	Christianity
76	Miscellaneous
85	Biology and Medicine
86	World Politics
93	Law
98	Media

Table 3. Names for the 15 Wikipedia topics.

shown are from the MPCA model that was built from the ctrl@listserv.aol.com mailing list. The queries were:

1. Keywords “art music” resulting 3095 emails between 01/1999–03/2003.

2. Single email address resulting 1766 emails between 01/1999–03/2003.

3. Thread initiated by the same email address as in query 2 resulting 232 emails between 02/99–10/02.

The topical strength of the query results and the strength of the topic are not comparable due to different normalization but the general trends are the interesting features here. It is worth noting that the temporal behavior of search results in Figures 4 and 5 are very similar. That is natural as the the constraints are an email address and a thread initiated by that email address respectively.

All the examples presented illustrate combined usage of different facets. The MPCA topics provide coarse summarization of the corpus. Topical trends per se give some insight of temporal behavior of a mailing list. However they are most useful when combined with search functionality, making possible to examine trends with respect to various subsets of the corpus by constraining the search with time, text or user. Analyzing topical trends against a well-known external model makes possible to quickly gain an overall idea about contents of an unseen corpus.

5 Conclusions

Multi-faceted nature of email data is a challenge for information retrieval. It provides also an interesting opportunity to experiment with integration of various methods, as presented in this paper. The possibility to chain the facets together makes the system efficient and usable even with large archives. Further work includes enhancements for each facet employed in the system. In addition we aim at building a polished demonstration system around the mailing list corpus.

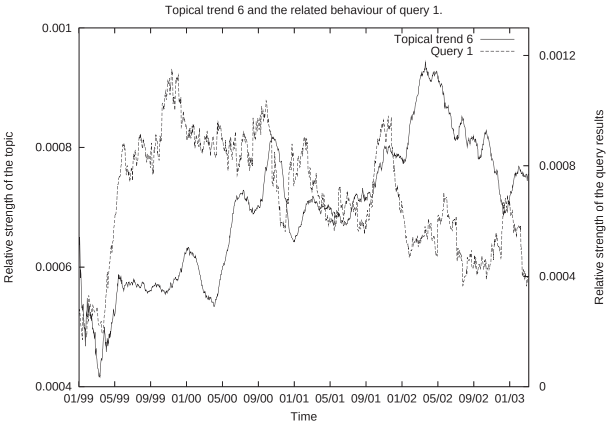

Figure 3. Query 1 and topic 6

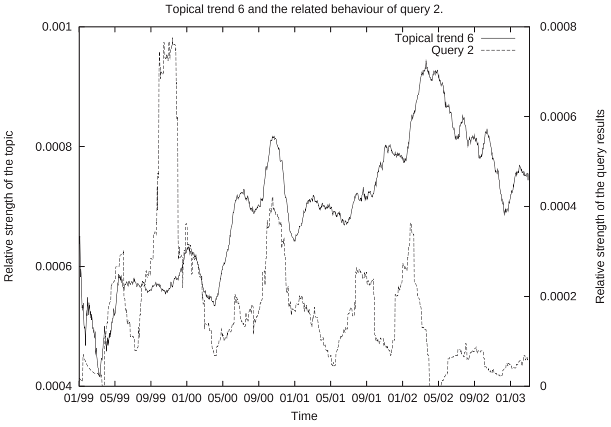

Figure 4. Query 2 and topic 6

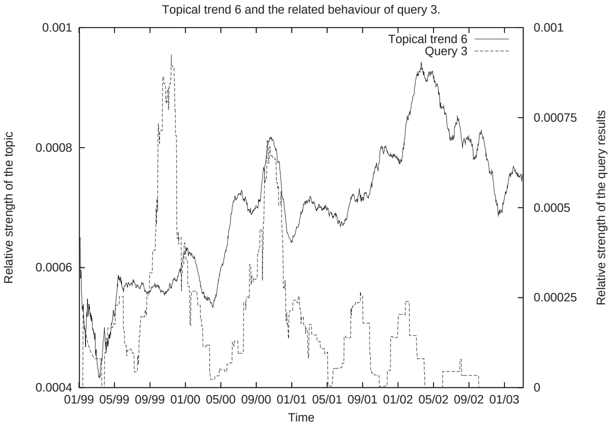

Figure 5. Query 3 and topic 6

Acknowledgements

This work was supported in part by the National Technology Agency, under the project Search-Ina-Box and by the Academy of Finland, under the project Prose. Further support was received by the IST Program of the European Community under the Alvis project. The authors express their gratitude to Jeff Breidenbach, administrator of the Mail Archive, for providing the data for this research.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[2] W. Buntine, J. Löfström, J. Perkiö, S. Perttu, V. Poroshin, T. Silander, H. Tirri, A. Tuominen, and V. Tuulos. A scalable topic-based open source search engine. In *International conference on web intelligence, WI2004*, pages 226–234. IEEE Computer Society, 2004.

[3] W. Buntine and S.Perttu. Is multinomial pca multi-faceted clustering or dimensionality reduction? In C. Bishop and B. Frey, editors, *Proc. of the Ninth International Workshop on Artificial Intelligence and Statistics*, 2003.

[4] W. B. Cavnar and J. M. Trenkle. N-gram-based text categorization. In *Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval*, pages 61–175, 1994.

[5] B. Dom, I. Eiron, A. Cozzi, and Y. Zhang. Graph-based ranking algorithms for e-mail expertise analysis. In *DMKD '03: Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 42–48. ACM Press, 2003.

[6] T. Griffiths and M. Steyvers. Finding scientific topics. *PNAS Colloquium*, 2004.

[7] B. Kerr and E. Wilcox. Designing remail: reinventing the email client through innovation and integration. In *CHI '04: Extended abstracts of the 2004 conference on Human factors and computing systems*, pages 837–852. ACM Press, 2004.

[8] J. Perkiö, W. Buntine, and S. Perttu. Exploring independent trends in a topic-based search engine. In *International conference on web intelligence, WI2004*, pages 664–668. IEEE Computer Society, 2004.

[9] J. Perkiö, W. Buntine, and H. Tirri. A temporally adaptive content-based relevance ranking algorithm. In *28th Annual Intl. ACM SIGIR Conference*, 2005.

[10] M. Porter. An algorithm for suffix stripping. In *Program*, volume 14, pages 130–137, 1980.

[11] A. C. Ron Bekkerman and A. McCallum. Extracting social networks and contact information from email and the web. In *Proceedings of CEAS 2004*, 2004.

[12] R. B. Segal and J. O. Kephart. Mailcat: an intelligent assistant for organizing e-mail. In *AGENTS '99: Proceedings of the third annual conference on Autonomous Agents*, pages 276–282. ACM Press, 1999.

[13] V. Tuulos and T. Silander. Language pragmatics, contexts and a search engine. In *International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, 2005.

[14] V. Tuulos and H. Tirri. Combining topic models and social networks for chat data mining. In *WI '04: Proceedings of the Web Intelligence, IEEE/WIC/ACM International Conference on (WI'04)*, pages 206–213. IEEE Computer Society, 2004.