Supporting Collaboration by Large Scale Email Analysis*

Michal Laclavík[1], Martin Šeleng[1], Marek Ciglan[1], Ladislav Hluchý[1]

Institute of Informatics, Slovak Academy of Sciences,
Dúbravská cesta 9, Bratislava, 845 07
laclavik.ui@savba.sk

Abstract. Email has become the most widespread Internet application. It is a tool supporting not only communication but also cooperation, task management, archiving, or information and knowledge management. Furthermore Email is a source of information on personal, job or community network of an individual or an organization. Email communication analysis allows extraction of social networks with further connection to people, organizations, locations, topics or time. This paper presents simple examples explaining how to extract a social network from email communication using approaches from the field of information retrieval and semantic annotation. We describe an experiment on the Hadoop distributed architecture designed to process large email archives.

Key words: email, social network, information extraction, metadata

1 Introduction

Mailing lists as well as common Email communication is a tool supporting cooperation. In addition it carries valuable information and knowledge that may be used to increase quality of cooperation support or information management. Email communication contains social networks of people communicating together and connected via their fields, projects or problems. Currently, there are no effective solutions that would focus on using information and knowledge extracted from archives or mailing lists.

Processing of Email communication using both information extraction and semantic annotation can create a social network graph with reference to other objects included in email communication. Such social network together with other metadata can improve information retrieval and collaboration.

At the same time email archives are becoming bigger and bigger both for personal and corporate email archives; therefore information extraction requires application of the methods that are scalable and able to process large data sources.

2 MapReduce

We have successfully ported semantic annotation into Grid [8] with good results, but porting of application, data management and results integration was not easy and time consuming task. Thus we have focused also on different parallel and distributed architectures as well. MapReduce [1] architecture developed by Google was used

* This work is supported by projects Commius FP7-213876, AIIA APVV-0216-07, SEMCO-WS APVV-0391-06, VEGA 2/7098/27.

with success on information retrieval tasks. Information extraction and pattern based annotation use similar methods such as information retrieval.
Google's MapReduce [1] architecture seems to be a good choice for several reasons:

- Information processing tasks can benefit from parallel and distributed architecture with simply programming of Map and Reduce methods
- Architecture can process Terabytes of data on PC clusters with handling failures
- Most of information retrieval and information extraction tasks can be ported into MapReduce architecture, similar to pattern based annotation algorithms. E.g. distributed *grep*[2] using regular expressions, one of basic examples for MapReduce, is similar to Ontea [5] pattern approach using regular expressions as well.

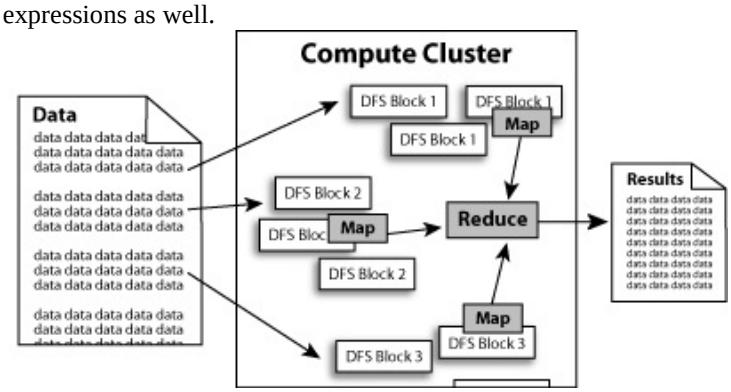

Fig. 1. MapReduce Architecture figure (source: Hadoop website).

On Figure 1 we can see main components of the MapReduce architecture: Map and Reduce methods, data in distributed file system (DFS), inputs and outputs. Several replicas of data are created on different nodes, when data are copied to DFS. Map tasks are executed on the nodes where data are available. Results of Map tasks are key value pairs which are reduced to results produced by Reduce method. All developer need to do, is implement Map and Reduce method. The architecture will take care of distribution, execution of tasks as well as fault tolerance. For more details on MapReduce please see [1].
Several open source implementation of MapReduce are available:

- Hadoop [2], developed as Apache project with relation to Lucene and Nutch information retrieval systems, implemented in Java. Hadoop is well tested on many nodes. Yahoo! is currently running Hadoop on 10,000 nodes [9] in production environment [6].
- Phoenix [10], developed at Stanford University, implemented in C++.
- Disco [11], currently an open-source implementation of the Map-Reduce. It was started at Nokia Research Center as a lightweight framework for rapid scripting of distributed data processing tasks. Disco is written in Erlang. Users of Disco typically write jobs in Python.

[2] Grep is a flexible search-and-replace function that can search one or more files for specified characters and/or strings

We are using the Ontea system [5] for information extraction. Ontea is written in Java, therefore we decided to use Hadoop MapReduce implementation.

3 Extraction of Metadata and Social Network

Using MapReduce and its Hadoop [2] implementation we have conducted several experiments with email archives. In these experiments we have extracted a social network as a directed and valued graph that can be used for search or information retrieval from emails [3], as a recommendation system, for example Acoma [7] (see Fig. 2) or used for information management similarly as in the system Xobni [4].

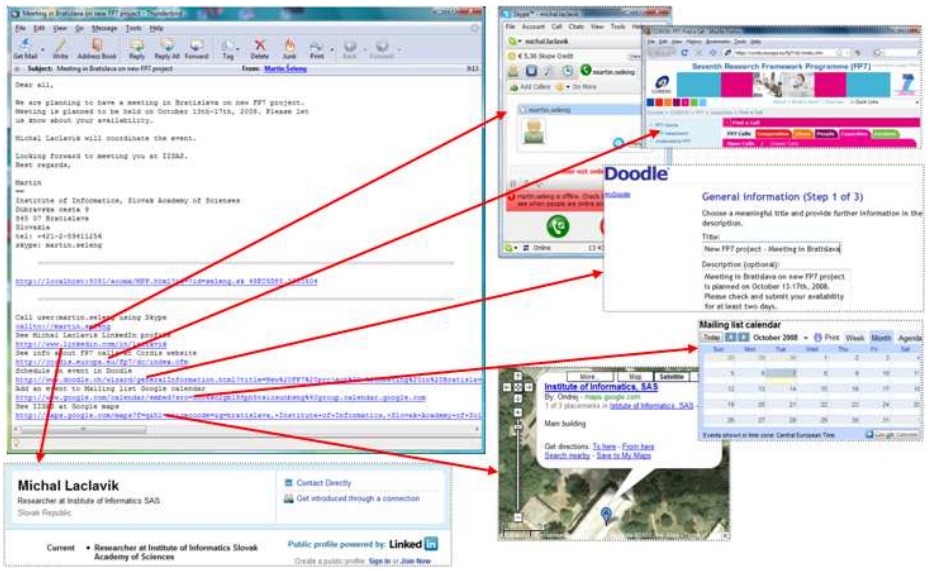

Fig. 2. Use of extracted metadata in recommendation system Acoma. Acoma process email communication and adds suggestions for actions in form of hints with links in the context of the email message. Hints context is detected using information extraction and semantic annotation. Context can contain detected objects such as communication people, organizations or their interaction networks, locations, projects, contact information or various application specific business objects. See [7] for more details on Acoma.

Using semantic annotation we have transformed graph nodes represented by email addresses to other nodes representing people (email addresses grouping) – Fig. 4, groups, organizations or countries – Fig. 3. We have employed the method of a pattern search in order to perform information extraction and semantic annotation Ontea [5].

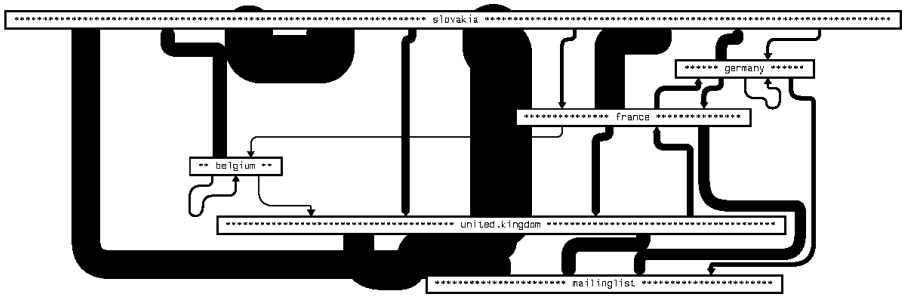

Fig. 3. Social network of communicating countries based on identical social network as on the Fig. 4, transformed into communicating countries using a semantic model.

Within extraction and semantic analysis generated the following semantic metadata:

- A social network of communicating people
- Amount of interaction (the number of sent and received emails)
- Transformation of a social network to another network; of organizations or people for instance.
- The graph can show other metadata on objects included in email communication with dependence on an application semantic model such as an organization, a company or geographical locations.

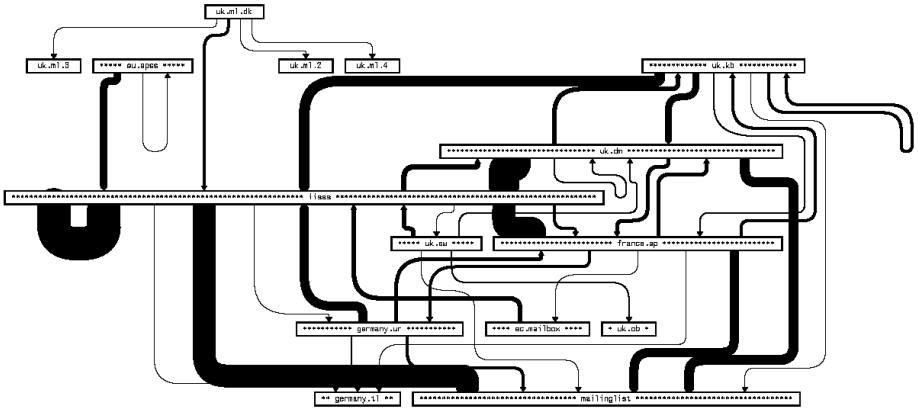

Fig. 4. Social network of communication within cooperation with partners, while the Institute of Informatics' Personnel are linked using semantic transformation as one graph node.

Within our previous paper [5] we tested semantic annotation and information extraction at the Hadoop cluster illustrating application of the Hadoop solution for such tasks. Social network extraction from email was another task we ported to the Hadoop cluster. Experiments were carried out on two different Hadoop clusters:

- Cluster #1: a 6-node (12 cores) Hadoop cluster based on Intel(R) Core(TM)2 CPU 2.40GHz with 4GB RAM hardware on all nodes.
- Cluster #2: a 8-node (32 cores) Hadoop cluster based on Intel(R) Core(TM)2 Quad CPU Q9550 2.83GHz 4GB RAM hardware on all nodes.

We have analyzed email archives of different sizes. Size of the biggest analyzed email archive was 1.8 GB. From the data we have extracted 1420 graph vertices/nodes. Results of the analyzes can be seen in the Table 1.

Table 1. Experiment on Hadoop cluster

Cluster #	Cores	Speed up	Time
1	2	-	51 min 33 sec
1	12	5.95	8 min 40 sec
2	32	14.87	3 min 28 sec

Table 1 shows that proposed approach is scalable and benefit from execution in the Hadoop environment. We do not provide execution time on single machine since the algorithm was written directly for Hadoop as Map and Reduce methods but we provide time of execution on Hadoop Cluster #1 with 1 node (2 cores) in the first row of the Table 1. If we compare speed up by adding the cores in the cluster. This is almost linear increase of performance.

Email archives processing on Hadoop architecture can scale to Terabytes of data which was proved also by Yahoo!, which use Hadoop in production environment [6] on web data. Advantage of Map Reduce architecture for social network analysis is also in natural way of integrating number of communication among people by Reduce method.

4 Conclusion

In our work we have shown how social network and related metadata can be extracted from large scale email archives. This metadata can be transformed and integrated with semantic model of application and used for improving information management, search or collaboration. Personal email social network was already used in systems such as Xobni [4]. In our approach we try to use it also on organizational or community level, where large email archives need to be processed and application specific semantics need to be used to detect the context of the message and to relate or transform extracted social network.

In our future work we would like to use extracted social network to improve recommendation tool Acoma [7] and thus provide tool for improving email based collaboration by exploiting personal, organizational or community social network extracted from the email communication.

References

1. Dean J., Ghemawat S.: MapReduce: Simplified Data Processing on Large Clusters, Google, Inc. OSDI'04, San Francisco, CA (2004)
2. Lucene-hadoop Wiki, HadoopMapReduce, http://wiki.apache.org/lucene-hadoop/HadoopMapReduce (2008)
3. Einat Minkov, Ramnath Balasubramanyan, William W. Cohen: Activity-centric Search in Email; in CEAS 2008 also in Enhanced Messaging Workshop, AAAI 2008, http://www.cs.cmu.edu/~einat/activitySearch.pdf

4. MIT Technology Review: A New Look for Outlook - Xobni makes it easier to find relevant information buried in your inbox; http://www.technologyreview.com/Biztech/19463/?a=f, 2007
5. Michal Laclavik, Martin Seleng, Ladislav Hluchy: Towards Large Scale Semantic Annotation Built on MapReduce Architecture; In Proceedings of ICCS 2008; M. Bubak et al. (Eds.): ICCS 2008, Part III, LNCS 5103, pp. 331-338, 2008
6. Yahoo! Launches World's Largest Hadoop Production Application, Yahoo! Developer Network, http://developer.yahoo.com/blogs/hadoop/2008/02/yahoo-worlds-largest-production-hadoop.html, (2008)
7. Michal Laclavik, Martin Seleng, Emil Gatial, Ladislav Hluchy: Future Email Services and Applications; Proceedings of the Poster and Demonstration Paper Track of the 1st Future Internet Symposium (FIS'08), CEUR-WS, ISSN 1613-0073, Vol-399, pages 33-35, 2008.
8. Michal Laclavik, Marek Ciglan, Martin Seleng, Ladislav Hluchy: Empowering automatic semantic annotation in grid. In Paralell processing and applied mathematics : 7th International Conference, PPAM 2007. - Springer-Verlag, 2008. ISBN 978-3-540-68105-2. ISSN 0302-9743, p. 302-311.
9. Open Source Distributed Computing: Yahoo's Hadoop Support, Developer Network blog, http://developer.yahoo.net/blog/archives/2007/07/yahoo-hadoop.html, (2007)
10. The Phoenix system for MapReduce programming. http://csl.stanford. edu/~christos/sw/phoenix/. (2008)
11. Disco Web site. http://discoproject.org/, (2008)