Text Mining and Qualitative Analysis of an IT History Interview Collection

Petri Paju[1], Eric Malmi[2], and Timo Honkela[2]

[1] Cultural History Department, University of Turku, Finland
petpaju@utu.fi
[2] Department of Information and Computer Science
Aalto University School of Science and Technology, Finland
eric.malmi@gmail.com; timo.honkela@tkk.fi

Abstract. In this paper, we explore the possibility of applying a text mining method on a large qualitative source material concerning the history of information technology in one nation. This data was collected in the Swedish documentation project "From Computing Machines to IT." We apply text mining on the interview transcripts of this Swedish documentation project. Specifically, we seek to group the interviews according to their central themes and affinities and pinpoint the most relevant interviews for specific research questions. In addition, we search for interpersonal links between the interviews. We apply a method called the "self-organizing map" that can be used to create a similarity diagram of the interviews. We then discuss the results in several contexts including the possible future uses of text mining in researching history.

Keywords: Documentation project, IBM, IT-history, methods development, self-organizing map (SOM), Sweden, text mining

1 Introduction

Research on the history of computing has, for a considerable time, involved developing methods, especially the application of oral history methods for conducting interviews, as Thomas Misa wrote in the book *History of Nordic computing 2* [1]. However, until recently, the field has shown limited interest in developing methods of its own or in using computational methods. Notwithstanding, new experiences from a couple of projects has brought attention to using the internet and other electronic means in *creating* sources for research. This contrasts with the tradition of collecting source material from the archives or publications but is closer to the established practices of interviewing for research purposes. One of these recent projects is the Swedish documentation project "Från matematikmaskin till IT" (http://ithistoria.se/), in English "From Computing Machines to IT," which during 2007–2008 had produced a large database concerning the Swedish history of information technology. This collection of information includes among other things (more than) 160 interviews and 50 organized group discussions, or witness seminars, almost all available as transcripts in the Internet [2]. They are in Swedish language, and therefore understandable in most of the Nordic area. Nevertheless, because of its formidable size, the "data" is difficult to handle as a whole and especially so to a non-Swede or non-expert who is unfamiliar with the many details of the subjects, people and (national) topics.

In this paper, we explore the possibility of applying text mining on the interview transcripts of the Swedish documentation project. Specifically, we seek to group the interviews according to their central themes and affinities and pinpoint the most relevant interviews for specific research questions. In addition, we have searched for links between the interviews and we use those links to outline interpersonal connections of the group of interviewees.

However, our study includes seventy-four interviews so it does not cover all of the interviews conducted in the project. The choice of the interviews was based on the selection of transcribed texts at the beginning of this study.

We apply a method called the "self-organizing map" [3] that can be used to create a similarity diagram of the interviews [4]. In the following, we shortly introduce the method, present the analysis results and discuss them in the contexts of the history of computing, the Nordic countries and as a methodological tool for future challenges with the masses of historical electronic sources.

2 Methods

The self-organizing map (SOM) algorithm [3] has been developed for the analysis and visualization of large masses of complex data. It projects data non-linearly on to a two-dimensional plane in such a way that the original structure of the data is retained as well as possible. The outcome of the SOM analysis is a map in which entities, such as people, words, sentences, or documents, are positioned on the map according to similarity with respect to some property. The SOM serves several analytical functions. First, it provides a mapping from a high dimensional space into a low-dimensional space, thus providing a suitable means for analysis and visualization of complex data. Second, the SOM reveals topological structure of the data. The topological distance between two points in the map is proportional to the distance between the points in the original input space.

The SOM algorithm originally grew out of early neural network models, especially models of associative memory and adaptive learning [5]. The underlying motivation was to explain the spatial organization of the brain's functions, as observed especially in the cerebral cortex. Nonetheless, the SOM was not the first step in that direction. The spatially ordered line detectors of von der Malsburg [6] and the neural field model of Amari [7] preceded the development of the self-organizing map. However, the self-organizing power of these early models was rather weak. The crucial invention of Kohonen was to introduce a system model that is composed of at least two interacting subsystems of different nature. One of these subsystems is a competitive neural network that implements the winner-take-all function, but there is also another subsystem that is controlled by the neural network and which modifies the local synaptic plasticity of the neurons in learning [8]. The learning is restricted spatially to the local neighborhood of the most active neurons. Only by means of the separation of the neural signal transfer and the plasticity control has it become possible to implement an effective and robust self-organizing system [9].

From a present-day point of view, the SOM, as an unsupervised statistical machine learning method, compares to classical unsupervised quantitative research methods such as multidimensional scaling or clustering. The SOM has been extensively used to analyze numerical data in many areas, including various branches of industry, medicine, and economics and the number of references to SOM-based research articles is currently over eight thousand [10–12]. This popularity of the SOM in a large area of scientific disciplines has raised Kohonen to be (one of) the most cited Finnish scientists in any field. The use of the SOM has also been extended into the analysis of text data (see, e.g., [4, 13]). It can be used for the study of large amounts of material such as e-mails, web sites, interview transcripts, etc.

Janasik et al. have shown that the SOM-based text mining process improves the quality of the inferences drawn by researchers doing qualitative research [14]. The SOM specifies a holistic conceptual space. The meaning of some item in an analysis is not based on a predefined definition but is the emergent result of a number of encounters in which the item is used in some context. Moreover, the emergent prototypes on the map are not isolated instances (as in many forms of cluster analysis), but they influence each other in the adaptive formation process [14]. In its emphasis on the grounded nature of knowledge, the SOM approach aligns well with the central epistemological presuppositions of traditional and revised grounded theory (see e.g. [15]). In grounded theory, the concepts and conclusions are drawn from the research material rather than determined beforehand as hypotheses to be tested. This is also the approach taken commonly in modern historical research.

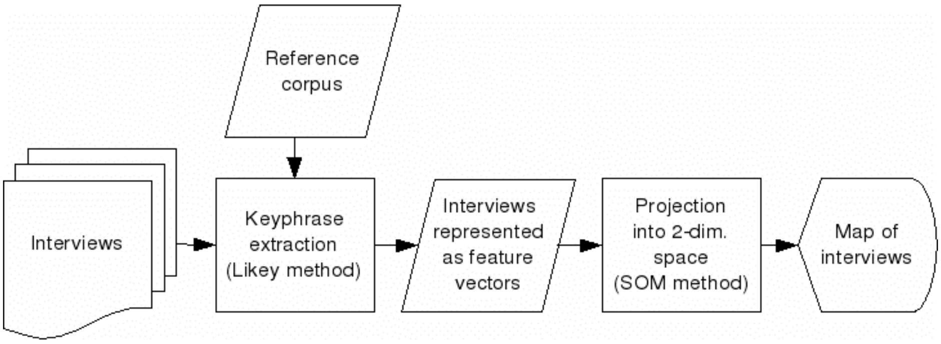
Reference corpus
Interviews
Keyphrase extraction (Likey method)
Interviews represented as feature vectors
Projection into 2-dim. space (SOM method)
Map of interviews

Fig. 1. The basic processes in creating map of interviews.

In our analysis, the interviewees are automatically positioned on the SOM according to the central themes discussed in their interviews (see Fig. 1 as an illustration of the overall process).

The themes are automatically extracted from the texts using language-independent keyphrase extraction (Likey) method [16]. Likey utilizes a reference corpus to highlight keyphrases that occur frequently in an interview compared with the reference corpus. Likey approach to extracting keyphrases is based on observations about the statistical properties of natural language. We use the distribution of phrase frequencies in texts to determine the significance of a phrase in a document by comparing it to the corresponding frequency in a baseline reference. In a related method introduced by Damerau, terms are ranked according to the likelihood ratio and the top *m* terms are used as index terms [17]. Likey produces keyphrases using relative ranks of *n*-gram frequencies. It is a simple language-independent method: the only language-specific component is a reference corpus in the corresponding language. Likey does not utilize any commonly used language-dependent (pre)processing such as stemming, stop word lists, part-of-speech tagging or syntactic parsing [16]. The availability of the software implementing the method may be requested from one of the authors (TH).

In the present analysis, we use a concatenation of all the interviews as the reference corpus for the Likey method. The effect of using Likey is that we are reasonably independent of researchers' biases. On the other hand, the choice of the reference corpus is a subjective matter. Choosing the whole corpus of interviews could be a rather neutral alternative but it has some specific effects. For instance, it does not select the acronym "IBM" as a keyword by the method as it is common in this corpus in general. If some other corpus, such as Europarl (the extracted proceedings of the EU Parliament), were used a word like "IBM" would end up in the list of keyphrases.

When the keyphrases have been selected (this can also be conducted manually instead of using the Likey method), a term-document matrix can be formed. In the matrix, each row corresponds to a document (interview), and each term (keyphrase) is represented by a column. The cells in the matrix then contain the number of occurrences of each term in each document. We may normalize the matrix by dividing the number in each cell by the row sum. This operation makes sure that documents of different length are considered in the analysis in a balanced manner. This matrix can then be given as input to some statistical software package that contains the SOM algorithm. In our case, we have used Matlab software, specifically the SOM Toolbox developed for Matlab (available for free at http://www.cis.hut.fi/projects/somtoolbox/). One can find more detailed information on using SOM-based text mining for qualitative research in [14].

In addition to the central themes, we also extract the references between the interviewees. Specifically, a directed link from interviewee A to interviewee B is drawn if A mentions B's surname in the interview. Thus, we get a social network of the interviewees drawn from top of the SOM.

3 Results

The resulting SOM is presented in Fig. 2. The interviewees that have had common themes discussed in their interviews are located close to each other. In a color picture, the lighter the area between two people, the more in common their interviews have.

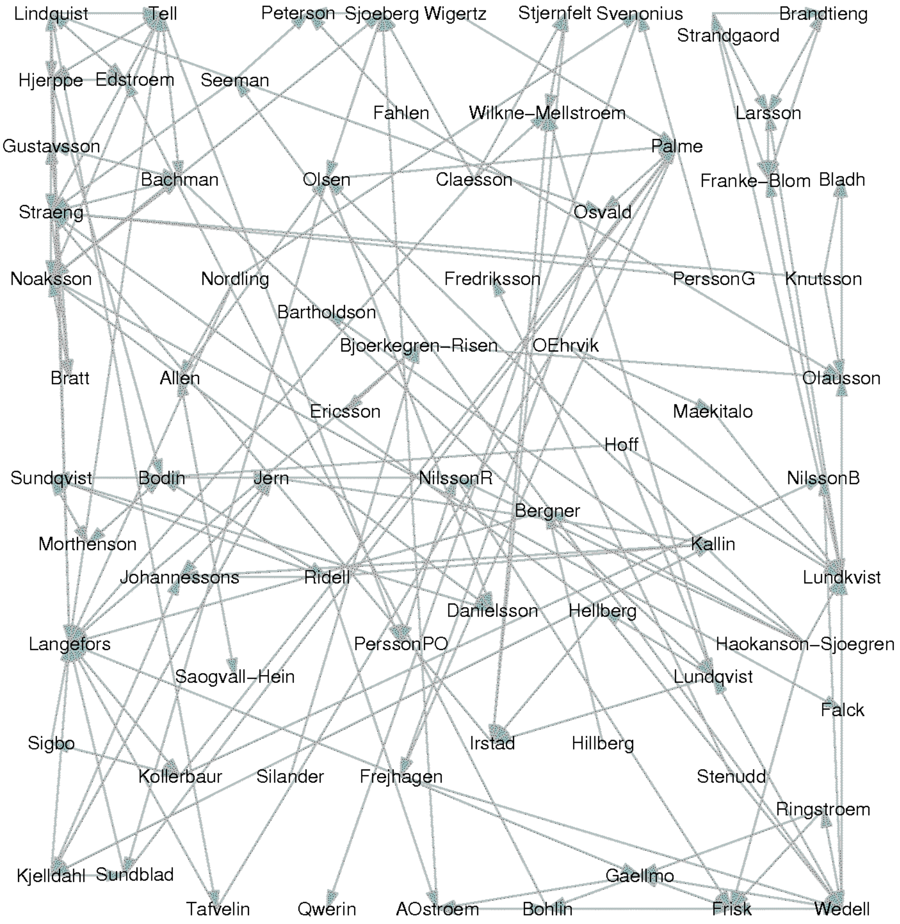

Fig. 2. A self-organizing map of interviewees based on interviews produced in the Swedish documentation project "Från matematikmaskin till IT." The interviewees that have had common themes discussed in their interviews are located close to each other. Also references between the interviewees are indicated: a directed link from a person to another is drawn if the first mentions the surname of the other in his or her interview.

To visualize different perspectives of this SOM, a terminology distribution diagram for each central theme is created (in short: theme map). The terminology distribution diagrams of the most popular themes are presented in Fig. 3 and Fig. 4. A dark area on a terminology distribution diagram implicates that on the SOM the interviewees within the same area have discussed the theme indicated in the title of the sub-diagram. For example, from the diagram on "numerisk analysis" (numerical analysis) one can see that the interviewees on the bottom-left corner of the SOM have mentioned numerical analysis in their interviews. Furthermore, some of these interviewees have also been discussing human-computer interaction (Fig. 4: “människa-datorinteraktion”).

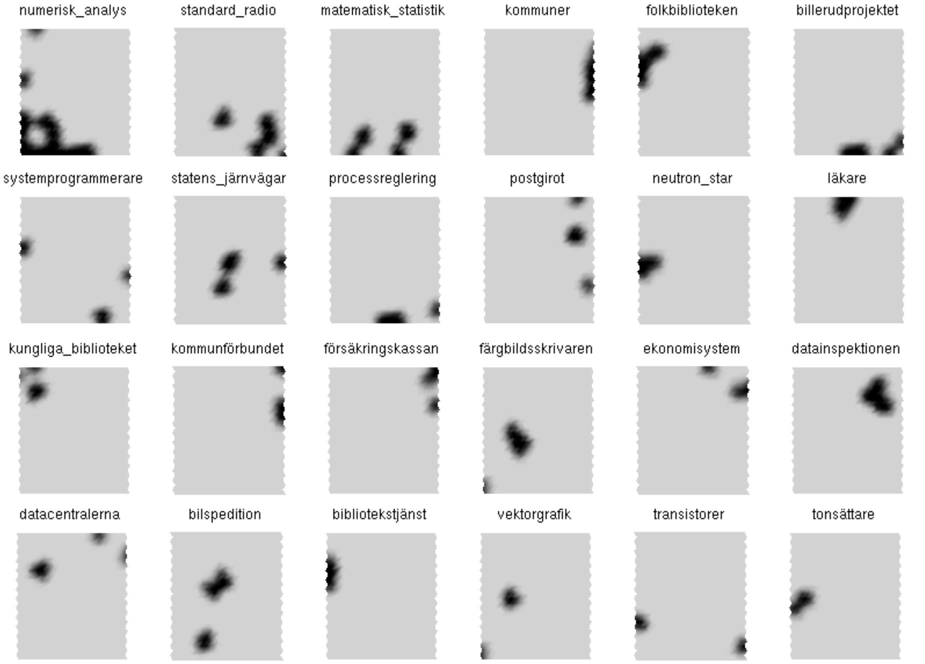

Fig. 3. A terminology distribution diagram of twenty-four themes commonly discussed in the interviews. Areas with dark color in each sub-diagram indicate that the theme has been handled in the interviews that have been mapped in the corresponding area on the SOM. For instance, the lower right corner of the sub-diagram “billerudprojektet” is dark which coincides with the corner where the surnames Wedell and Frisk are located (see Fig. 2). To learn how to interpret the theme maps imagining them on top of the SOM map (Fig. 2), see Fig. 5 (on the Billerud project).

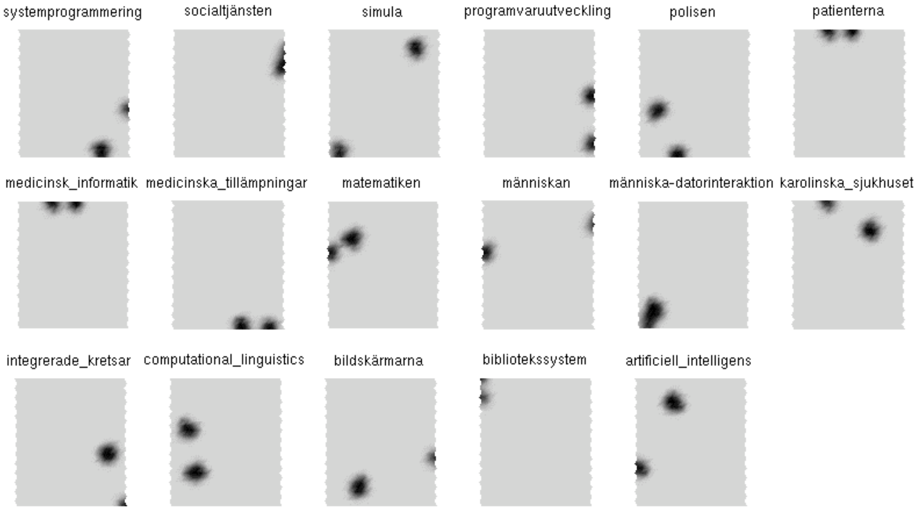

Fig. 4. A terminology distribution diagram of another seventeen themes commonly discussed in the interviews (see the caption of Fig. 3 for more details).

Comparing the theme area of the human-computer interaction with the SOM map (Fig. 2), we can see that the interviewees (Lars) Kjelldahl and (Anita) Kollerbaur and perhaps their neighboring interviewees discussed this topic. After looking at other theme area maps, we notice these two have also talked about “Simula” (the programming language), “vektorgrafik” (vector graphics), and to a lesser extent “färgbildsskrivaren” (color printer) as well as the already mentioned numerical analysis.

The terminology distribution diagrams thus allow us to find the interviewees who have mentioned certain themes. It has to be emphasized that each diagram illuminates a particular aspect of the same map. Thus, a particular area on each diagram, for instance the upper right corner, always refers to the same items, in this case particular persons and their interviews. The theme maps can also be used for finding correlations between the themes. There is, e.g., an intuitive connection between the planes “patienterna” (patients) and “medicinsk information” (medical information). In addition, some less obvious connections can also be found, such as:

“försäkringskassan” (social insurance office) –
“kommunförbundet” (association of local governments)
and
“ekonomisystem” (accounting system) –
“datacentralerna” (computer centers).

A total of 188 interpersonal references were found from the interviews. These connections are visualized with arrows (ties) and they include 35 mutual ties. Majority of the ties are short suggesting that two interviewees, one of whom has mentioned the other, are likely to have discussed similar topics as well. In other words, the structure of the map based on the contents of the interviews coincides well with the structure of the social network. From yet another point of view, one can say that people who have similar interests also tend to know each other. If the same number, i.e. 188, interpersonal references were created randomly between the interviewees, the average length of the ties would be considerably longer than in this real case, and the network diagram would appear much more complex.

The social network also brings out the interviewees with many references. Three interviewees with the most references are Börje Langefors (13 references), Björn Tell (7), and Gunnar Wedell (7).

4 Discussion and Conclusions

In this section, the above results will be discussed in the contexts of the history of computing, especially in the Nordic countries and as a methodological tool for future challenges with the masses of historical electronic sources.

As a test case towards qualitative analysis, we have utilized the SOM maps especially in studying IBM in Sweden. With the help of, for instance, keywords provided by the editors of the interviews, we identified the bottom right hand corner as having a high concentration of IBM related people and topics. From the theme area maps we see that “billerudprojektet” (Billerud project), “systemprogrammerare” (systems programming), “processreglering” (process control), and some other themes come up in that section of the SOM. A closer look at the interviews indicated that this corner has gathered interviews about the IBM Nordic Laboratory, established in 1960 near Stockholm and clearly one focus point of the documentation project. Further, reading the close-by interviews we can quickly find other related interviews, such as the ones dealing with the ALGOL compiler project of the IBM laboratory. Especially Bengt Gällmo and Birgitta Frejhagen talked about the ALGOL project. Another job of the IBM laboratory was a process control project in Billerud Company, depicted in one of the theme maps (see also Fig. 5). We thus get a good chance of finding most of the relevant information on such details promptly. However, the big picture of, for instance IBM’s influence in Sweden or effects on the interviewees’ careers, is much more complex and requires close reading of several of the interviews in the SOM’s ‘IBM corner’.

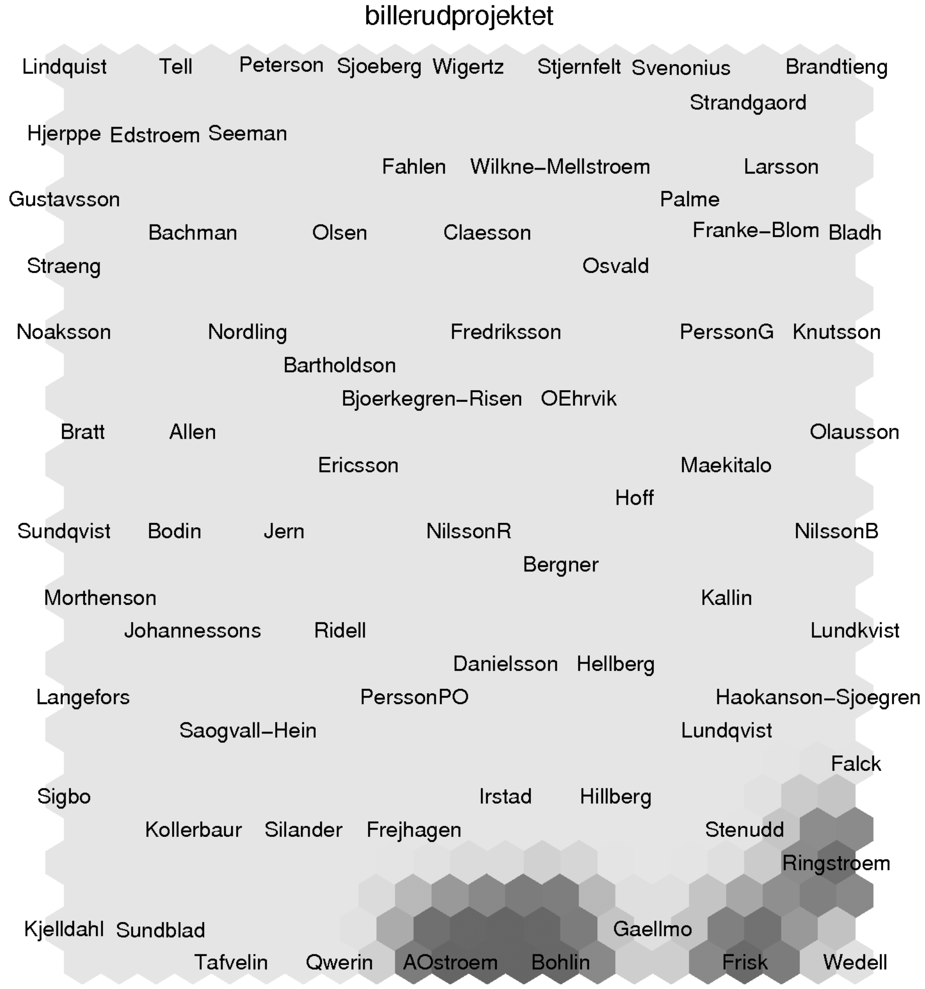

Fig. 5. An enlarged terminology distribution diagram or theme map shown on top of the SOM map indicates which interviewees talked about the Billerud project. They were, from right to left, (Karl Johan) Åström, (Torsten) Bohlin, (Tage) Frisk, and (Ingemar) Ringström.

Since IBM was originally not chosen as a possible SOM theme because of its very high frequency in the data, we wanted to see how common it really is in this material. We decided to compare the words IBM and Ericsson to see their spread in the SOM data.

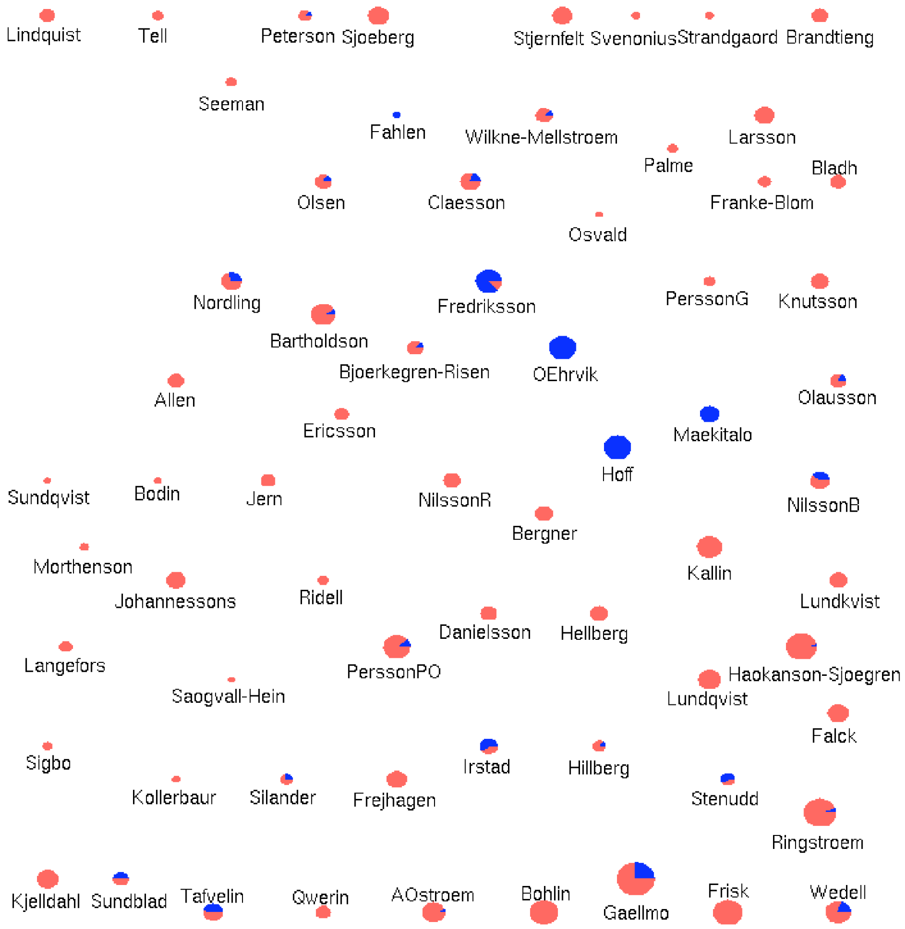

Fig. 6. Word frequencies of ‘IBM’ (light) and ‘Ericsson’ (dark) in the interviews.

This map (see Fig. 6) reveals that IBM, unsurprisingly perhaps, is very commonly referred to in the interviews of the documentation project. The mapping also indicates an Ericsson concentration (OEhvik, Hoff) that could be interesting.

We hope to have shown that the SOM method can contribute to analyzing data also in history research and furthermore, can be useful for a special field such as history of computing in the Nordic countries. Especially when the Swedish language is applied, as in this documentation project under analysis, the data is understandable in most of the Nordic area. The SOM makes the data of even such a formidable size, and rich with countless details of subjects, people and (nationally specific) topics, accessible for a non-Swede and/or for a non-expert user.

In the future, we suggest using the SOM method for historical data (also in English) and as a new way of producing user guidelines for large information archives such as in the field of history of computing and those in the Charles Babbage Institute at the University of Minnesota. Moreover, with ever-increasing amount of electronic source materials becoming objects of historical scrutiny in the future, we foresee many uses for text mining with the SOM method in organizing the analysis of such electronic source collections.

Many historians oriented towards qualitative research may find it challenging to use a method such as terminology extraction or the SOM, with which they potentially do not have any prior experience. For the moment, before easy-to-use implementations of the necessary software are available, we recommend that these kinds of studies be conducted in collaborative and interdisciplinary contexts because this would ensure that various aspects of the necessary methodological expertise are available. This is particularly true for automatic terminology extraction and some advanced uses of the SOM [14].

Acknowledgments. We warmly thank Prof. Reijo (Shosta) Sulonen who with his vision brought the authors of this paper together. His discussions with Prof. Yrjö Neuvo originally generated the basic ideas behind this study. We are also grateful to Academician Teuvo Kohonen whose work, including the development of the self-organizing map algorithm, has been foundational to a whole research area.

References

1. Misa, T. J.: Organizing the History of Computing: 'Lessons Learned' at the Charles Babbage Institute. In: Impagliazzo, J., Järvi, T., Paju, P. (eds.) History of Nordic Computing 2. IFIP AICT 303. Springer, Berlin, 1–12 (2009)
2. Lundin. P.: Documenting the Use of Computers in Swedish Society between 1950 and 1980: Final Report on the Project "From Computing Machines to IT." Working Papers from the Division of History of Science and Technology, TRITA/HST 2009/1, Stockholm (2009)
3. Kohonen. T.: Self-Organizing Maps. Third edition. Springer, Berlin (2001)
4. Honkela, T., Kaski, S., Lagus, K., Kohonen, T.: WEBSOM – Self-organizing maps of document collections. Proceedings of WSOM'97, Workshop on self-organizing maps (pp. 310–315). Helsinki University of Technology, Espoo, Finland (1997)
5. Kohonen T.: Self-Organization and Associative Memory. Springer, Berlin (1984)
6. von der Malsburg, C.: Self-organization of orientation sensitive cells in the striate cortex. Kybernetik, 14: 85–100 (1973)
7. Amari, S.: Topographic organization of nerve fields. Bulletin of Mathematical Biology, 42: 339–364 (1980)
8. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43:59–69 (1982)
9. Kohonen, T., Honkela, T.: Kohonen network. Scholarpedia, 2(1): 1568 (2007)
10. Kaski, S., Kangas, J., Kohonen, T.: Bibliography of self-organizing map (SOM) papers: 1981-1997. Neural Computing Surveys, 1: 102–350 (1998)
11. Oja, M., Kaski, S. and Kohonen, T.: Bibliography of Self-Organizing Map (SOM) Papers: 1998–2001 Addendum. Neural Computing Surveys, 3: 1–156 (2003)
12. Pöllä, M., Honkela, T., Kohonen, T.: Bibliography of Self-Organizing Map (SOM) Papers: 2002–2005 Addendum. Technical Report TKK-ICS-R23, Helsinki University of Technology (2009)

13. Lagus, K., Kaski, S., Kohonen, T.: Mining massive document collections by the WEBSOM method. Information Sciences, 163(1–3), 135–156 (2004)
14. Janasik, N., Honkela, T., Bruun, H.: Text mining in qualitative research: Application of an unsupervised learning method. Organizational Research Methods, 12(3): 436–460 (2009)
15. Castellani, B., Castellani, J., Spray, S. L.: Grounded neural networking: Modeling complex quantitative data. Symbolic Interaction, 26(4): 577–589 (2003)
16. Paukkeri, M.S., Nieminen, I.T., Pöllä, M., Honkela, T.: A Language-Independent Approach to Keyphrase Extraction and Evaluation. In: Scott, D. and Uszkoreit, H. (eds.) COLING 2008, 22nd International Conference on Computational Linguistics, Posters Proceedings, 18–22 August 2008, Manchester, UK, 83–86 (2008)
17. Damerau, F.: Generating and evaluating domain-oriented multi-word terms from text. Information Processing and Management, 29(4): 433–447 (1993)