Analyzing an Email Collection Using Formal Concept Analysis

Richard Cole[1] and Peter Eklund[1]

School of Information Technology, Griffith University
PMB 50 GOLD COAST MC, QLD 9217, Australia
r.cole@gu.edu.au, p.eklund@gu.edu.au

Abstract. We demonstrate the use of a data analysis technique called formal concept analysis (FCA) to explore information stored in a set of email documents. The user extends a pre-defined taxonomy of classifiers, designed to extract information from email documents with her own specialized classifiers. The classifiers extract information both from (i) the email headers providing structured information such as the date received, from:, to: and cc: lists, (ii) the email body containing free English text, and (iii) conjunctions of the two sources.

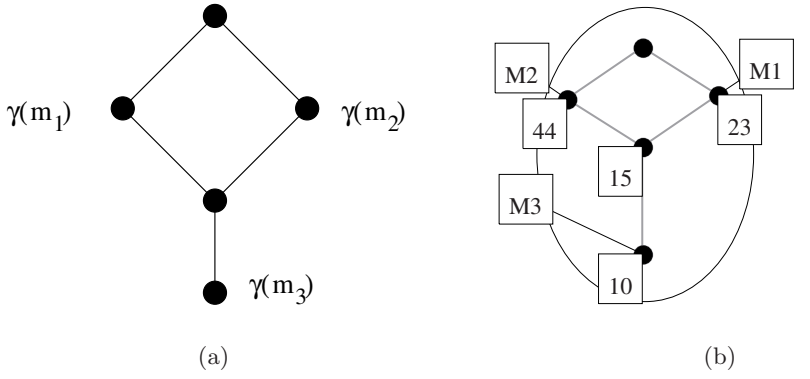
1 Formal Concept Analysis

Formal Concept Analysis (FCA) [8,3] is a mathematical framework for performing data analysis that has as its fundamental intuition the idea that a concept is described by its intent and its extent. FCA models the world as being composed of objects and attributes. The choice of what is an object and what is an attribute is dependent on the domain in which FCA is applied. Information about a domain is captured in a *formal* context, which is a triple $\mathbb{K} = (G, M, I)$ in which G is a set of objects, M is a set of attributes, and $I \subset G \times M$ is a relation saying which objects possess which attributes.

A *formal* concept is a pair (A, B) where A is a set of objects called the extent, and B is a set of attributes called the intent. A must be the largest set of objects for which each object in the set possesses all the attributes in B. The reverse must be true also of B. More precisely, a *formal* concept of the context (G, M, I) is a pair (A, B), with $A \subseteq G$, $B \subseteq M$, $A = \{a \in G \mid \forall\, b \in B\ (a, b) \in I\}$ and $B = \{b \in M \mid \forall\, a \in A\ (a, b) \in I\}$.

The fundamental theorem of FCA states that the set of *formal* concepts of a *formal* context forms a complete lattice. This complete lattice is called a *concept* lattice, and is usually denoted $B(\mathbb{K})$. A complete lattice is a special type of partial order in which the greatest lower bound and least upper bound of any subset of the elements in the lattice must exist. A lattice may be drawn via a line diagram (see Figures 1 and 4) [7]. For each attribute in m there is a maximal concept that has m in its extent. We shall use the function $\gamma : M \to B(\mathbb{K})$ to denote the mapping from attributes to their maximal concepts.

J.M. Żytkow and J. Rauch (Eds.): PKDD'99, LNAI 1704, pp. 309–315, 1999.



(a) (b)

Fig. 1. (a) A concept lattice with the implication $m_1 \wedge m_2 \rightarrow m_3$ (b) A concept lattice with the partial implication $Pr(M1|M2) = 15/44$.

A concept lattice, generally denoted $B(\mathbb{K})$, is representationally equivalent to the attribute logic that exists over the attributes in the context. For example (see Fig. 1(a)), the proposition $\forall g \in G\ p_1(g) \wedge p_2(g) \rightarrow p_3(g)$ where $p_i(g)$ means that object g has attribute m_i will be true, if and only if the greatest lower bound of $\gamma(m_1)$ and $\gamma(m_2)$ will be greater than or equal to $\gamma(m_3)$ in the concept lattice. Since this information is represented diagramatically it is more accessible than a list of conditional probabilities.

Labels on the lattice (see Fig. 1(b)) attached above each concept show the introduced intent[1] while the labels attached below show the size of the extent of the concept. It is possible to determine the intent of a concept by collecting all intent labels on an upward path from the concept. For example the concept labeled $M3$ in Fig. 1(b) also has $M2$ and $M1$ in its intent.

The concept lattice can also represents partial implications or conditional probabilities[5]. For example, if we wanted to know the probability that an object from G has m_1 given that it has $M2$, denoted $Pr(M1 \mid M2)$, this would be given by $|Ext(\gamma(M1) \wedge \gamma(M2))|/|Ext(\gamma(M2))| = 15/44$. The two numbers in this ratio being present in the diagram (see Fig. 1(b)).

2 Background

Previous applications of FCA may be divided into two categories. Those that generate a large concept lattice — the number of concepts is roughly the square of the number of documents — of all terms and documents and those that employ conceptual scaling.

Godin et. al [4] proposed navigating though a set of text documents via a large concept lattice. Each concept visited was presented in a window listing its intent, and the user moved to subsequent elements nodes via selection of additional terms. Carpineto and Romano [1] proposed navigation in a concept lattice of all terms using a fish-eye viewer.

[1] The introduced intent is that part of the intent that is not found in the intent of any more general concepts.

Wille and Rock [6] implemented a system for a library catalogue, using a software system called TOSCANA, in which a large number of sub-lattices were designed by a subject librarian. A visitor to the library could choose a "theme" previously defined by the librarian. This was seen as an advantage in the library environment since the users of the system are generally unfamiliar with reading lattice diagrams.

In a sense these approaches represent different extremes. In the first instance the user has a maximum number of choices when navigating, while in the second the user is presented with carefully constructed views of the data. The approach outlined in this paper attempts to strike a middle road allowing the user to construct and modify scales in response to learning information about the data. It is novel in that it allows the user to define a hierarchy over the search terms and presents the user with a dynamic environment for the creation and modification of scales by the user.

3 Hierarchy of Classifiers

The task of extracting information from text documents is a difficult one. The language used in email documents is often informal, makes extensive use of abbreviations, and is highly contextualized. For this reason we do not attempt to do any deep extraction of information from email texts but rather recognize key terms.

We experimented with classifiers that recognize regular expressions, either from email headers, or within the body of the email itself.

```
1  CLASSIFICATION "Email-Analysis"
2  ...
3  "Year 1994 - Sep:Nov" Date:
4  "^[A-Z][a-z]+, [0-9]+ (Sep|Nov)
5      1994" ; ...
6  "Melfyn mentions Barbagello"
7      From: melfyn Body: David ;
8
9  "Mentions Melfyn" Body: [mM]elfyn ;
10 ...
11 END CLASSIFICATION
```

(a)

```
1 BEGIN ORDER "Email-Analysis"
2 ...
3 "Melfyn mentions Barbagello"
4     < "Mentions Barbagello" ;
5
6 "From Melfyn" < "From DSTC" ;
7 "From DSTC"   < "DSTC" ;
8 ...
9 END ORDER
```

(b)

Fig. 2. (a) Classifiers: a file expressing classifiers for the terms of interest via regular expressions. (b) Hierarchy: a file expressing the hierarchical ordering of classifiers.

The example in Figure 2(a) shows a portion of the classifier file, generated by our taxonomy editor. Lines 3 and 4 show the definition of a classifier. The classifier recognizes the attribute whose name is "Year 1994 - Sep:Nov". It matches the date field of an email message with a regular expression that recognizes

dates between September and November of 1994. Lines 6 and 7 show a classifier that detects emails sent by "Melfyn" in which "David" is mentioned within the text of the email.

The classifiers associate attributes with emails and the result is stored in an inverted file index. A hierarchy is defined by a set of subsumption rules defined by the user. For instance, a portion of the example file produced by our taxonomy is presented in Fig. 2(b).

Line 3 introduces the implication that an email with "Melfyn mentions Barbagello" implies that the email "Mentions Barbagello".

Associated with each attribute is a primary name, e.g. "Melfyn mentions Barbagello", and a set of descriptions, for example "Melfyn mentions Barbagello" might also be recognized by "Melfyn Lloyd mentions David Barbagello". These extra descriptions are used in the next section in which the data is explored.

4 Conceptual Scaling

After defining a taxonomy of attributes which are associated with email documents by classifiers defined in the previous section, it is necessary for the user to choose a small number of attributes, usually less than 10 [2]. The user does this with a specific question in mind and with the aid of the program depicted in Fig. 3.

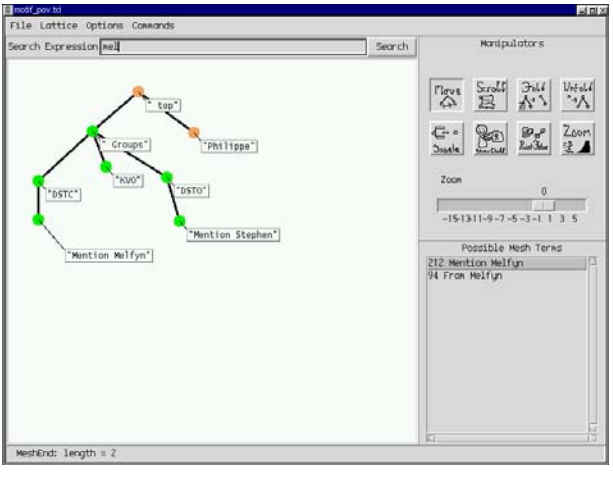


Fig. 3. Conceptual Scale Creation Tool

The user is interested in the ways in which the attributes combine in the email collection. For example she might be interested in emails "from Melfyn", "about the DTSC", and "mentioning Barbagello". She searches for appropriate attributes based either (i) on their description or (ii) their location in the hierarchy.

To locate an attribute via its description, the user may enter a text search. For instance "Melfyn" would match with all attributes having a description containing a word with the prefix "Melfyn". Alternatively, the attribute "From Melfyn" might be located as an element immediately below "From DSTC Personnel". In either case, the results of the search operation are displayed in the lower right hand panel of the tool shown in Fig. 3 and clicking on the attribute adds it to the diagram.

The *conceptual scale* is a subset of the attributes selected by the user, and displayed in the left hand panel of the tool shown in Fig. 3. It is desirable for the user to gain an impression of how the attributes are related with respect to their taxonomical ordering defined in the previous section.

We represent the attributes using a Hasse diagram (see Figure 1(a)). Using a Hasse diagram to diagramatically represent the relative ordering of a subset of elements from an ordered set raises a number of questions. Should we preserve (i) the covering relation, (ii) the ordering relation, and (iii) meet and joins where they exist. Preserving the covering relation, while straight forward, is cumbersome since it produces long chains in the diagram and introduces a large number of extra elements.

Preserving the ordering relation by itself, would for many queries, produce a single anti-chain[2]. We preserve the ordering relation that we then close under join. This induces a new covering relation computed in response to updates to the diagram. The diagram is automatically drawn using a force directed placement heuristic that attempts to minimize change to the diagram. All changes to the diagram are animated to help the user preserve their mental map of the diagram.

The user can remove join irreducible elements — those not required by the join closure requirement — from the diagram. An attempt to remove a join reducible element results in user feedback showing the elements preventing its removal.

After the user has selected attributes to the required level of specificity. The scale may be used to construct a concept lattice showing the concepts generated by the emails and the attributes selected by the user.

5 Analyzing the Lattice

Fig. 4 shows the concept lattice resulting from the terms identified in the scaling process shown in Fig. 3. The diagram is expressed as a lattice product. Each of the small black circles represent a concept, all the concepts of the bottom large oval have "Philippe" in the intent. In other words these concepts refer to emails associated with Philippe via one or more of the classifiers defined in Fig. 2(a).

Compare the numeric labels of the top two concepts in each of the ovals. "793" is the number of emails associated with the term "Philippe" while "5022" is the total number of emails in the test set. We infer from this that 793/5022 or 15.7% of all emails are associated with "Philippe". Moving down from the top numeric label "5022" to its child "4108" (also labeled "Groups") reveals that 81% of all emails in the test set are associated with the term "Groups".

[2] An anti-chain is a set of elements that have no relative ordering.

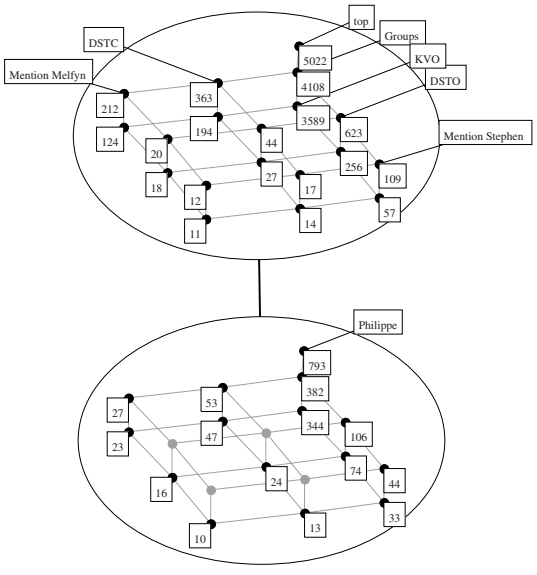


Fig. 4. Analysis of the relationships between the terms "Mention Stephen, "KVO", "DSTC", "DSTO", "Melfyn" and the target "Philippe".

Similarly, if we move down from the top label "793" in the lower oval to its child "382" we see that only 48% of emails associated with "Philippe" concern "Groups" — a possible reason being that much of the email traffic within the email set associated with "Philippe" is not group related.

"Groups" are further divided into group categories: these can be read from the top oval identified with the labels "KVO", "DSTC", and "DSTO". Note that "KVO" concerns the majority of group email traffic with 3589/4108 = 87.3%. Emails associated with the group labels "DSTO" and "DSTC" are read from the label "44" in the top oval. Finding the corresponding circle in the lower oval we notice that it is grey. This represents an implication. We move from the grey circle down through the lattice to the label "24". This point includes the extra attribute "KVO". The inference is that emails associated with "DSTC", "DSTO" and "Philippe" are all associated with the term "KVO".

The diagram can be viewed as a three dimensional space containing thematic planes. Consider the plane defined by the labels 109,17,12,11,14,57 in the upper oval of Fig 4. This plane represents the impact of the term "Mention Stephen" on the other named terms "KVO", "DSTC", "DSTO", "Melfyn". The plane is parallel to two other planes (those above it, to the right) and by considering corresponding points in each of these planes we can measure the influence of the term "Mention Stephen" on the way emails in the test set are partitioned by that term.

A more specific inference concerns the points labeled "11" and "12" in the upper oval. "12" is the number of emails associated with the combination of "Mention Stephen" and "Mention Melfyn". Therefore, 11/12 of these emails

also mention the term "KVO". The inference is the high correspondence of the use of "KVO" in the context of emails involving 'Melfyn' and 'Stephen'. Moving to the bottom circle in the lower oval labeled "10" we infer that 10/11 emails also mention "Philippe".

In summary less than half of the email associated with "Philippe" is group related (382/793). When "Philippe" is mention in the context of a group it mostly concerns the "KVO' group (344/382 = 90%). Emails associated with both the "DSTC" and "DSTO" groups that are "Philippe" related always mentioned the "KVO" group (24/24 — inferred via the grey circle). Finally, "Philippe" is mentioned in 24/44 (55%) of emails involving both the "DSTC" and "DSTC" but is mentioned in 83% of correspondence mentioning "Stephen" and "Melfyn".

We can draw the inference from the analysis of this email data that "Philippe" is the important factor of common interest between "Stephen" and "Melfyn". It is also clear from the email analysis that "Philippe" is a more important topic of discussion to the "KVO" group (344/382) than to the "DSTC" (53/382) and "DSTO" (106/382) generally.

6 Conclusions

This paper has described the use of a suite of tools designed to allow an investigation of data retrieved from email. The data is retrieved from the emails with the aid of a hierarchy of classifiers that extract useful terms and encode known implications. Further implications, both complete and partial, are then investigated by means of a nested line diagram.

References

1. C. Carpineto and G. Romano. A lattice conceptual clustering system and its application to browsing retrieval. *Machine Learning*, 24:95–122, 1996.
2. R. Cole and P. Eklund. Scalability of formal concept analysis. *Computational Intelligence*, 15(1):11–27, 1999.
3. B. Ganter and R. Wille. *Formal Concept Analysis: Logical Foundations.* Springer Verlag, 1999.
4. R. Godin, J. Gecsei, and C. Pichet. Design of a browsing interface for information retrieval. *SIG-IR*, pages 246–267, 1987.
5. Michael Luxenburger. Implications, dependencies and Galois drawings. Technical report, TH Darmstadt, 1993.
6. T. Rock and R. Wille. Ein TOSCANA—Erkundungssystem zur Literatursuche. In G. Stumme and R. Wille, editors, *Begriffliche Wissensverarbeitung. Methoden und Anwendungen*, Berlin–Heidelberg, 1997. Springer–Verlag.
7. Frank Vogt and Rudolf Wille. TOSCANA a graphical tool for analyzing and exploring data. In *Graph Drawing '94*, LNAI 894, pages 226–223. Springer Verlag, 1995.
8. Rudolf Wille. Concept lattices and conceptual knowledge systems. In *Semantic Networks in Artificial Intelligence.* Pergamon Press, Oxford, 1992. Also appeared in *Comp. & Math. with Applications,* 23(2-9), 1992, p. 493-515.