Mining Contrast Concepts Representing Differences

Gupei Qin Makoto Haraguchi Yoshiaki Okubo

Graduate School of Information Science and Technology
Hokkaido University

This paper is concerned with a problem of finding contrast concepts formalized in terms of Formal Concept Analysis (FCA). A contrast concept shows that we can observe a difference between a given pair of contexts (relations) in the sense that the concept has many sub-concepts in one context but less in the other. An algorithm for extracting every contrast concept is presented. We design a depth-first search algorithm with some pruning rules which can reduce our search space. In our experimentation with datasets of news articles, we veryfy effectiveness of our method.

1. Introduction

In this paper, we are concerned with a problem of finding contrast concepts. If someone is interested in studying the cultural differences between different places for a common topic with much attention according to local news reports, it is usually necessary for him/her to read all news at those places and compare them one by one. Obviously, the amount of news is huge and it is extremely difficult and time-consuming. In order to help this kind of task, we try to extract contrast concepts which represent differences between two databases. Our problem of detecting contrast concepts is formalized in terms of Formal Concept Analysis (FCA) [1]. The main task is to extract every concept which frequently appears in both of two concept lattices constructed from the databases and satisfies some constraints on the sub-lattices rooted with the concepts. The sub-lattice constraint requires that a contrast concept has many sub-concepts in one database but less those in the other database. That is, intuitively speaking, in the former, the sub-lattice rooted with the contrast concept is well organized, but in the latter poorly organized. For example, in the above news article example, this means that in the former, topics concerned with the contrast concept are fully discussed in various points of views. In the latter, on the other hand, topics w.r.t. the contrast concept are not yet mature. With the help of such a contrast concept, it is expected that we might be able to notice some valuable differences between those places.

In order to efficiently extract contrast concepts w.r.t. a given pair of databases, we design a depth-first search algorithm with some pruning rules. In general, since a concept lattice is organized with a huge number of concepts, we try to reduce the number of attributes by applying Spectral Clustering with an Extended K-Means. In our experimentation,we try to extract contrast concepts found for news articles in “Mainland” and “Hong Kong” of China.

Contact: Makoto Haraguchi
IST, Hokkaido University
N-14 W-9, Sapporo 060-0814, JAPAN
E-mail: mh@ist.hokudai.ac.jp

2. Preliminaries

In this section, we introduce some basic terminologies in Formal Concept Analysis [1].

Let $\mathcal{O}$ be a set of *objects* and $\mathcal{A}$ a set of *attributes*. For a binary relation $R \subseteq \mathcal{O} \times \mathcal{A}$, a triple $\langle \mathcal{O}, \mathcal{A}, R \rangle$ is called a *formal context*. If $(o, a) \in R$, we say that the object o has the attribute a or a is associated with o. It is easy to see that a transaction database in *Frequent Pattern Mining* [5] can be regarded as a formal context and vice versa.

Given a formal context $\langle \mathcal{O}, \mathcal{A}, R \rangle$, for a set of objects $X \subseteq \mathcal{O}$ and a set of attributes $Y \subseteq \mathcal{A}$, the *derivation operator* “$\prime$” is defines as

$$\begin{aligned} X' &= \{a \in \mathcal{A} \mid \forall o \in X, a \text{ is associted with } o\} \quad \text{and} \\ Y' &= \{o \in \mathcal{O} \mid \forall a \in Y, o \text{ has } a\}. \end{aligned}$$

The former computes the set of attributes shared by every object in X. The latter, on the other hand, returns the set of objects with Y.

Based on the operator, for a set of objects $X \subseteq \mathcal{O}$ and a set of attributes $Y \subseteq \mathcal{A}$, a pair of X and Y, (X, Y), is called a *formal concept* (or simply a concept) in the formal context if and only if $X' = Y$ and $Y' = X$, where X and Y are called the *extent* and the *intent* of the concept, respectively. From the definition, it is easy to see that $X'' = X$ and $Y'' = Y$. That is, a formal concept is defined as a pair of *closed* sets of objects and attributes under the derivation operator. The set of all formal concepts in a context $\mathcal{C}$ is denoted by $\mathcal{FC}(\mathcal{C})$.

Let $C_i = (X_i, Y_i)$ and $C_j = (X_j, Y_j)$ be concepts. If $X_i \subseteq X_j$ (or equivalently $Y_i \supseteq Y_j$), then we say C_i is a *sub-concept* of C_j or C_j is a *super-concept* of C_i and denote it by $C_i \preceq C_j$. Under the ordering, the set of formal concepts in a formal context forms a lattice, called a *concept lattice*.

3. Problem of Mining Contrast Concepts

In this section, we formalize our problem of mining contrast concepts w.r.t. a given pair of contexts. Before providing the formal definition, we first introduce some notions.

For a concept C in a context $\mathcal{C}$, the set of sub-concepts

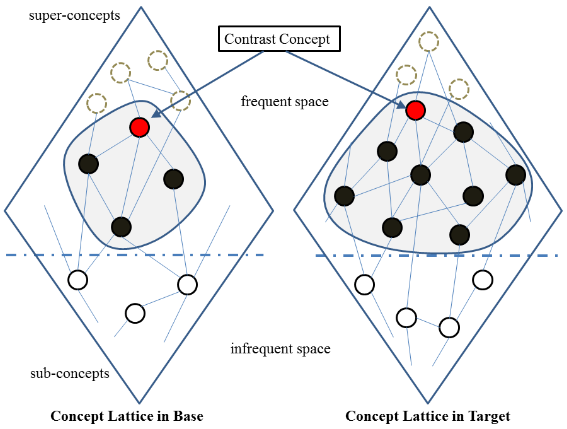


Figure 1: Contrast Concept

of C is denoted by $SUB_{\mathcal{C}}(C)$, that is,

$$SUB_{\mathcal{C}}(C) = \{D \in \mathcal{FC}(\mathcal{C}) \mid D \preceq C\}.$$

Let *min_sup* be a minimum support threshold. For a concept $C = (X, Y)$ in a context $\mathcal{C} = \langle \mathcal{O}, \mathcal{A}, R \rangle$, if $|X|/|\mathcal{O}| \geq$ *min_sup*, then C is said to be *frequent* in $\mathcal{C}$. The set of frequent concepts in $\mathcal{C}$ is denoted by $\mathcal{FC}_\delta(\mathcal{C})$.

3.1 Contrast Concepts

For a given pair of formal contexts, one is regarded as a *base*, denoted by $\mathcal{C}_B$, and the other as a *target*, denoted by $\mathcal{C}_T$. Intuitively speaking, a contrast concept w.r.t. the context pair is a formal concept C such that

- C is frequent in each of $\mathcal{C}_B$ and $\mathcal{C}_T$.
- The number of frequent sub-concepts of C in $\mathcal{C}_B$ is less than a threshold θ_B.
- Conversely, the number of frequent sub-concepts of C in $\mathcal{C}_T$ is greater than a threshold θ_T, where $\theta_T > \theta_B$.

Figure 1 illustrates a contrast concept we try to extract.

3.2 Contrast Concept Mining

We present here a formal definition of our problem, *Contrast Concept Mining.*

Given a pair of formal contexts $\mathcal{C}_B = \langle \mathcal{O}_B, \mathcal{A}_B, R_B \rangle$ and $\mathcal{C}_T = \langle \mathcal{O}_T, \mathcal{A}_T, R_T \rangle$, we assume $\mathcal{A}_B = \mathcal{A}_T$ so that for any concept C in one context, we can obtain a corresponding concept in the other which is semantically equivalent to C. More precisely speaking, for a concept $C_T = (X, Y)$ in $\mathcal{C}_T$, we consider a concept C_B in $\mathcal{C}_B$, defined as $C_B = (Y', Y'')$, to be equivalent to C_B.

Definition 1 (Contrast Concept Mining)
Let $\mathcal{C}_B$ and $\mathcal{C}_T$ be a base and a target contexts, respectively. Given δ, θ_B and θ_T, *Contrast Concept Mining* is a problem of finding every concept $C_T = (X, Y)$ in $\mathcal{C}_T$ satisfying the following constraints:

Constraint on Support:
$C_T \in \mathcal{FC}_\delta(\mathcal{C}_T)$ and $C_B = (Y', Y'') \in \mathcal{FC}_\delta(\mathcal{C}_B)$.

Constraint on Sub-Structure:
$|SUB_{\mathcal{C}_T}(C_T)| > \theta_T$ and $|SUB_{\mathcal{C}_B}(C_B)| < \theta_B$.

■

4. Algorithm for Extracting Contrast Concepts

In this section, we discuss how to extract contrast concepts w.r.t a given pair of base and target contexts, $\mathcal{C}_B$ and $\mathcal{C}_T$. We can design a depth-first algorithm with some pruning rules based on the following simple observations.

Observation 1
If a concept C in a context $\mathcal{C}$ is not frequent, then any concept C' in $\mathcal{C}$ such that $C \preceq C'$ is also not frequent. ■

It is equivalent to a property well-known as *anti-monotonicity of support* in Frequent Pattern Mining. It is obvious that if a concept C is not frequent, we do not need to examine every sub-concept of C because it can never be a contrast concept due to its infrequentness.

Observation 2
For a concept C, if $|SUB_{\mathcal{C}}(C)| < \delta$ holds, then for any concept C' such that $C \preceq C'$, we always have $|SUB_{\mathcal{C}}(C')| < \delta$. ■

From the property, it is easy to see that if a (frequent) concept C in $\mathcal{C}_T$ has less than δ_T sub-concepts, then we do not need to examine every sub-concept of C because it can never have a sufficient number of frequent sub-concepts.

With the help of the pruning rules, we can extract our contrast concepts in depth-first manner. More concretely speaking, as a basic procedure, for a contrast concept $C_i = (X_i, Y_i)$, the intent Y_i is expanded by adding an attribute $y \in \mathcal{A}_T \setminus Y_i$ and then computing an immediate sub-concept of C, $C_{i+1} = (X_{i+1} = (Y_i \cup \{y\})', Y_{i+1} = (Y_i \cup \{y\})'')$, in $\mathcal{C}_T$.

If C_{i+1} is not frequent or does not have enough sub-concepts, that is $|Y_{i+1}| < \delta$ or $|SUB_{\mathcal{C}_T}(C_{i+1}) < \delta_T$, then we discard C_{i+1} and try to examine another immediate sub-concept of C by expanding Y_i with a different $y \in \mathcal{A}_T \setminus Y_i$ as backtrack.

If C_{i+1} is frequent and has enough sub-concepts in $\mathcal{C}_T$, we then check whether the corresponding concept in base has less than θ_B sub-concepts in $\mathcal{C}_\mathcal{B}$. If it is true, C_{i+1} is output as a contrast concept and then tried to further expand with an attribute $y \in \mathcal{A}_T \setminus Y_{i+1}$. Otherwise, we just try to expand C_{i+1} because we could find some contrast concepts as sub-concepts of C_{i+1}.

With the initial concept $(\mathcal{O}_T, \mathcal{O}'_T)$, our expansion process is recursively iterated in depth-first manner until no concept remains to be examined.

Remarks: Reducing Complexity of Concept Lattice

In general, the concept lattice constructed from a given formal context has a huge number of concepts. In some practical cases, therefore, it would be necessary to reduce complexity of concept lattice, preserving the semantics of the original context.

As is presented below, we verify effectiveness of our method for a pair of contexts obtained from news articles. In those original articles, many individual terms are closely related or have a similar conceptual meaning. For example, "Chicago Bulls", "Los Angeles Lakers" and "New York Knicks" are closely related to "NBA" and have the same conceptual meaning, "Basketball Team in NBA". In such a case, it would be reasonable to identify them into an abstract term. As a result, the original context (an article-term relation) can be compressed into a smaller context with a less number of (abstract) terms. Furthermore, some infrequent concepts with small extents in the original context might be combined into a concept which is possibly frequent in the compressed context. Thus, with this kind of compression, we can expect to reduce complexity of the original concept lattice.

5. Experimental Results

In order to verify effectiveness of our method, we have implemented a system based on the method and made some experimentations. The system has been coded in Java and run on a PC with Intel$^{\circledR}$ Core$^{\text{TM}}$-i3 (2.93 GHz) processor and 2GB main memory.

5.1 Dataset

We have collected Mainland and Hong Kong news stream text articles for one month during the period from October 1st to October 31st, 2012. In total, $11,732$ Mainland news articles have been obtained from *Xinhua News Agency* [*1], while $13,609$ Hong Kong news articles from two newspaper companies, *Oriental Daily News* [*2] and *The Sun* [*3]. After a standard pre-processing, a morphological analysis and a term extraction based on TF-IDF values, we have obtained $20,542$ terms for the former and $23,248$ for the latter. Then we have obtained an article-term relation from each article set as a context.

The numbers of terms in the contexts seem to be a little bit large. In order to compress the original contexts, we have applied Spectral Clustering [3] with an extended k-means in which the number of clusters is automatically adjusted based on inner variance of clusters during clustering process. Based on the clustering result, we have obtained $16,272$ and $19,030$ terms, respectively.

The context based on the articles in Hong Kong is set to the base and that based on Mainland articles to the target.

5.2 Extracted Contrast Concepts

Under the parameter setting of $\delta = 0.001$ and $\theta_B = \theta_T = 10$, we have extracted contrast concepts. Some example contrast concepts are shown in Figure 2

For the concept $\{\mathit{Obama}\}$, the size of the extent in each context is almost the same. However, the number of sub-concepts in $\mathcal{C}_T$ (Mainland articles) is roughly twice that in $\mathcal{C}_T$ (Hong Kong articles). Sub-concepts in Mainland articles are, for example, $\{\mathit{Obama}, \mathit{debate}\}$, $\{\mathit{Obama}, \mathit{Romney}\}$, $\{\mathit{Obama}, \mathit{China}\}$, $\{\mathit{Obama}, \mathit{vote}\}$, $\{\mathit{Obama}, \mathit{Japan}\}$, $\{\mathit{Obama}, \mathit{Iran}, \mathit{Israel}\}$, $\{\mathit{Obama}, \mathit{Syria}\}$, $\{\mathit{Obama}, \mathit{unemploymentrate}\}$ and so on. On the other hand, sub-concepts in Hong Kong articles are $\{\mathit{Obamadebate}\}$, $\{\mathit{Obama}, \mathit{Romney}\}$, $\{\mathit{Obama}, \mathit{China}\}$, $\{\mathit{Obama}, \mathit{economic}\}$, $\{\mathit{Obama}, \mathit{U.S.dollar}\}$, $\{\mathit{Obama}, \mathit{Renminbi}\}$ and so on. From those sub-concepts, we can see both Mainland and Hong Kong are concerned with U.S. presidential election and the relation between the United States and China. However, we can also find some different view points. Mainland is more interested in politics than in Hong Kong. On the contrary, Hong Kong is more concerned with economy than in Mainland. Thus, it is expected that our contrast concepts have an ability to tell us some interesting differences between mainland and Hong Kong.

5.3 Effectiveness of Pruning Rules

In our computation, we can enjoy two pruning rules based on constraints on frequentness of concepts and the number of sub-concepts. Some experimental results have showed the former is actually effective in improving our computational efficiency. For the latter, on the other hand, we have observed little practical improvement on efficiency to be noted. Because in order to apply the pruning rule to a concept C, we need to count the number of frequent sub-concepts of C. However, due to the hugeness of the concept lattice, the task of counting sub-concepts is a little bit costly and spoils the pruning effect in our search.

5.4 Effectiveness of Compression

We have verified effectiveness of compression. Our experimental results have showed that after the compression, computation times can be reduced roughly in half. For example, for a pair of contexts with $10,000$ objects (news articles), we have spent about 8 minutes to obtain every contrast concepts with compression, while it has taken about 15 minutes without compression.

6. Conclusions

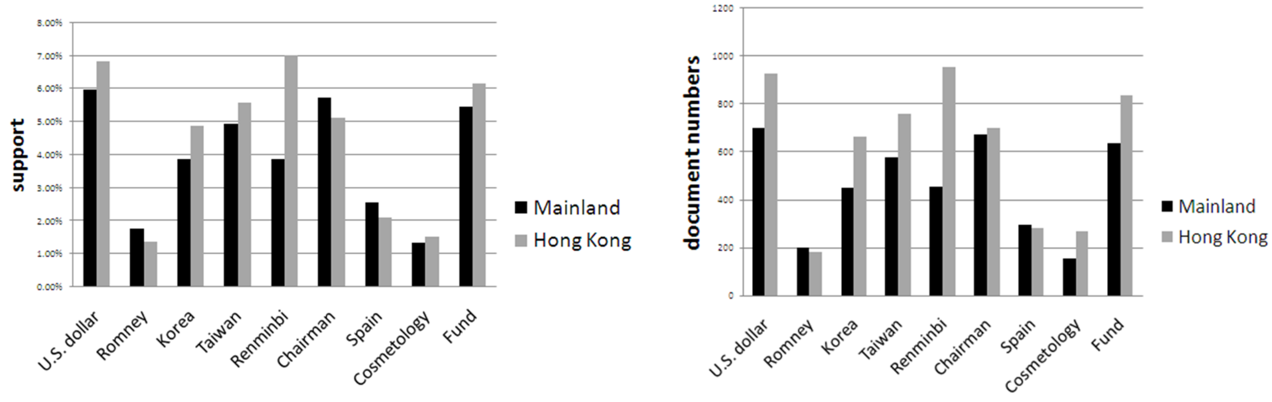
In this paper, we have discussed how to extract contrast concepts which represent differences between two formal contexts (relations). We have formalized our problem in terms of Formal Concept Analysis and designed a depth-first search algorithm with some pruning rules for efficient computation. In addition to the pruning rules, we have verified that compressing the original contexts is effective in improving efficiency of our computation. Our experimental results have showed that the proposed method is effective in finding interesting contrast concepts.

As future work, we try to extract contrast concepts from news articles in different countries. Moreover, it would be interesting to investigate contrast concepts in more than two databases. It would also be worth analyzing an adequate compression ratio which can provide us efficient computations without considerably loosing semantics of the original contexts.

[*1] http://www.xinhuanet.com/

[*2] http://orientaldaily.on.cc/

[*3] http://the-sun.on.cc



Intent of Concepts

Intent of Concepts

Figure 2: Example of Contrast Concepts

References

[1] B. Ganter and R. Wille, Formal Concept Analysis - Mathematical Foundations, Springer, 1999.

[2] V. Vychodil, A New Algorithm for Computing Formal Concepts, Proc. of EMCSR'08, pp. 15 - 21, 2008.

[3] A. Y. Ng, M. I. Jordan and Y. Weiss, On Spectral Clustering: Analysis and an Algorithm, Proc. of NIPS01, pp. 849 - 856, 2001.

[4] S. D. Bay and M. J. Pazzani, Detecting Group Differences: Mining Contrast Sets, Data Mining and Knowledge Discovery, 5(3), pp. 213 - 246, Kluwer Academic Publishers, 2001.

[5] J. Han, H. Cheng, D. Xin and X. Yan, Frequent Pattern Mining - Current Status and Future Directions, Data Mining and Knowledge Discovery, 15(1), pp. 55 - 86, Kluwer Academic Publishers, 2007.