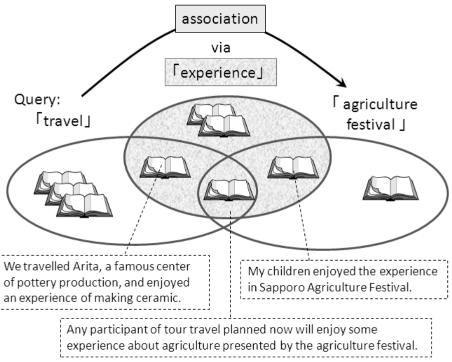
A Framework of Associative Search by Mediators

Hongjie Zhai[*1] Makoto Haraguchi[*1] Yoshiaki Okubo[*1]
Kiyota Hashimoto[*2] Sachio Hirokawa[*3]

[*1] Graduate School of Information Science and Technology, Hokkaido University
[*2] College of Sustainable System Sciences, Osaka Prefecture University
[*3] Research Institute for Information Technology, Kyushu University

In this report, we formalize a problem of associative search in terms of formal concepts. We first define a notion of aspects as intermediate bridge concepts connecting initial concepts and target concepts. The target concepts, drifted versions of the initial ones, must be conditionally similar to the initial ones with respect to the bride concepts. In order to find such target concepts effectively, we present a search procedure accessing two kinds of incident relations. One relation describes a person-feature relation, and contributes for suggesting potential persons related to the bridge concepts. The second type relation defines standard concepts of document-feature relations based on which the target concepts are derived under a similarity constraint given by the intents of bridge concepts.



We travelled Arita, a famous center of pottery production, and enjoyed an experience of making ceramic.

My children enjoyed the experience in Sapporo Agriculture Festival.

Any participant of tour travel planned now will enjoy some experience about agriculture presented by the agriculture festival.

Figure 1: An example of Association of Concepts

1. Introduction

In the studies of Information Retrieval, "Associative Search" [1] is sometimes used to shift queries during the search processes. The purpose of shifting queries is to suggest feature terms and documents not being properly expressed by the initial query or attracting user's interests. Fig. 1 illustrates an example of association, where ovals denote document sets defined by the corresponding set of feature terms, and query is also a set of terms. By the initial query, some documents with all the terms in it may involve another feature term, "experience" for instance, that can be an aspect of "travel" and therefore may guide us to another document set with the aspect. In this example, it might be "agriculture festival". From a viewpoint of formal concepts [5] of documents as objects and feature terms as attributes, the association from the concept of "travel" to "agriculture festival" is performed via the intermediate concept "experience". The association thus defined provides us a direct way to find semantic relationship among feature terms based on the information the documents carry, not using any kind of prior knowledge of feature terms.

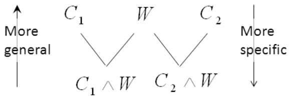


Figure 2: The diagram named W, where W is the intermediate concept connecting two concepts, C_1 and C_2

We have already proposed in [2] a similar transformation of concepts depicted as a concept diagram in Fig. 2. The relationship between the initial concept C_1 and its related C_2 via the intermediate W is schematically the same as [2]. The difference is found in the constraint setting. In [2] particularly, C_2 farer from C_1 is considered more preferable from the standpoint of discovering unexpected concepts sharing the same aspect W. As an extreme case, C_1 and C_2 may have no objects in common. To the contrary, we here require C_2 to be closer to C_1, given the intermediate W. We thus regard *conditional similarity* between C_1 and C_2 w.r.t. W, where the similarity must be based on the shared objects (documents).

The difficulty of associating concepts with other ones is however caused by the followings:

Selection of Candidate W (SCW): Every concept has several aspects, W in the W-diagram. So we have to choose some W from candidate concepts.

Selection of Target Concept (STC): Even when we select some adequate W, there may exist several C_2 with W as a shared aspect.

As STC problem is constrained by the conditional similarity given W, we can solve it by a constrained miner enumerating C_2 satisfying the constraint. On the other hand, SCW is more critical, because every association depends on the selection of W as an aspect. As long as we do not use

Contact: Hongjie Zhai and Makoto Haraguchi
IST, Hokkaido University
N-14 W-9, Sapporo 060-0814, JAPAN
E-mail:{zhaihj,makoto}@kb.ist.hokudai.ac.jp

any prior knowledge to restrict possible aspects, we have to use additional information and constraint relevant to SCW.

As such information, we here propose to use person-feature relationship defining additional concept lattice of concepts whose objects and attributes are persons and feature terms, respectively. Thus, we provide two incident relations, person-term relation and document-term relation, and use the former relation to take possible intermediate concepts W as bridges for connecting concepts. Intuitively, documents are associated with persons who write them. So we get to have features charactering the documents written by them. This defines a person-term relationship. Then, we introduce the following heuristics to control the SCW problem: *"When we wonder which direction of association is better to be developed, we normally ask some persons who have good knowledge of various areas of documents and feature terms. We call here such persons* mediators. *As more number of topics (as clusters of feature terms) they are concerned with, we regard them as better mediators."*

Using a membership function of fuzzy k-means clustering [3], we define a probability distribution on topics showing how a person is related to the topics, and then calculate its entropy as the degree of mediator. Then, every person is ranked according to the entropy, and the relevance to the initial query is tested by user who gives it. Some persons may be inadequate from the user's viewpoint, and another may be good conversely. We assign negative or positive signs to the former and the latter types of mediators, respectively. Then the target intermediate concepts are required to cover positive mediators and not to cover negative ones. Then the possible intermediate W is placed in a sublattice with initial concept as its top and the least common generalization of positive mediators as its bottom. This makes it possible for our search procedure to work in the restricted sublattice meeting with user's intention.

2. Person-Term Relation and Document-Term Relation

Let P be a set of *persons* and V a set of *terms* or *words* as a vocabulary. A *document* is written at the vocabulary V by a person in P. For a document d, the set of terms appeared in d is referred to as $terms(d)$ and the person by whom d is written as $person(d)$.

For a document set D, we can define a document-term relation $R_D \subseteq D \times V$, where $(d,t) \in R_D$ iff $t \in terms(d)$. It is often represented as a formal context $\mathcal{D} = (D, V, R_D)$.

In addition, a person-term relation $R_P \subseteq P \times V$ can also be defined from D. For the document set D, the set of documents written by a person $p \in P$ is denoted by D_p. Then, for a person p, the set of terms used by p in some document is given by $T_p = \cup_{d \in D_p} terms(d)$. Based on T_p, we can define $R_P \subseteq P \times V$ as $(p,t) \in R_P$ iff $t \in T_p$. It corresponds to a formal context $\mathcal{P} = (P, V, R_P)$.

3. Ranking Persons

If a person is related to various kinds of topics, it seems possible to access to several topics via the person. In this sense, it would be reasonable to consider that such a person is a good mediator for associative search and given a higher *mediator level.* To find a good mediator, assuming a cluster of terms (words) to be a *topic*, we evaluate a user's mediator level based on how closely related to topics the user is.

For a person-term context $\mathcal{P} = (P, V, R_P)$ and a document-term context $\mathcal{D} = (D, V, R_D)$, our mediator ranking of the persons in P is computed as follows:

1) $\mathcal{P}$ is projected by preserving all terms with $TF\text{-}IDF$ values greater than $(\max + \min) * \alpha$, where α is a control parameter, and max and min are the maximum and minimum values of $TF\text{-}IDF$. The projected context is denoted by $\tilde{\mathcal{P}} = (P, \tilde{V}, \tilde{R_P})$.

2) Based on $\tilde{V} \subseteq V$, the document context is also projected. The projected context is denoted by $\tilde{\mathcal{D}}$ and defined as $\tilde{\mathcal{D}} = (D, \tilde{V}, \tilde{R_D} = (R_D \cap (D \times \tilde{V})))$. Then we consider its corresponding matrix $M_{\tilde{\mathcal{D}}} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{|\tilde{V}|})$, where $\boldsymbol{w_i}^T = (d_{i1}, \ldots, d_{i|\mathcal{D}|})$ and for each j, if $(j,i) \in \tilde{R_D}$, then $d_{ij} = 1$ and othewise $d_{ij} = 0$.

3) The vectors $\boldsymbol{w_i}$ $(1 \leq i \leq |\tilde{V}|)$ are clustered into several groups of terms. Then each cluster is regarded as a topic and is represented by its central vector.

4) For the (projected) person context $\tilde{\mathcal{P}}$, we consider its corresponding matrix $M_{\tilde{\mathcal{P}}} = (\boldsymbol{p}_1^T, \ldots, \boldsymbol{p}_{|P|}^T)^T$, where $\boldsymbol{p_i}^T = (d_{i1}, \ldots, d_{i|\mathcal{D}|})$ and for each j, if $(i,j) \in \tilde{R_P}$, then $d_{ij} = 1$ and othewise $d_{ij} = 0$.

5) To transform each vector $\boldsymbol{p_i}^T$ into those in the document space, we apply *Singular Value Decomposition* to $M_{\tilde{\mathcal{D}}}$.

6) Degree of relatedness of a person p to a topic t is given by the distance measure used in fuzzy K-means in [3], $R(p,t) = \frac{1}{\sum_{t_i \in \mathcal{T}} (\frac{dist(p,t)}{dist(p,t_i)})^{\frac{2}{m-1}}}$ where $\mathcal{T}$ is the set of topics, m a parameter for fuzzy level and $dist(p_i, t_j)$ the Euclidean distance in the document space between p_i and t_j.

7) Regarding the vector of $R(p, t_i)$-value for each topic t_i as a probability distribution, the mediator level of p is given as the entropy of the distribution.

8) By sorting the persons in descending order of mediator levels, we define a ranked list of persons in P.

For each person p, the rank and the mediator level of p are referred to as $rank(p)$ and $level(p)$, respectively.

For each cluster i and j, we believe that $< c_i, c_j > \to 0$ by using the center of similar words. If one user has a more general interest, he/she will get involved into more separated topics. Entropy will help us to judge the variance of a user's interest.

In our clustering step, we use *Laplacian Eigenmaps* [6] to get closely related terms into one cluster. Moreover, for each topic, the document distribution is assumed to be Gaussian and the gaussian kernel is used to make this points linear. Our clustering algorithm is performed as follows:

1. Calculate the Matrix A using

$$A_{ij} = \{^{exp(-disi,j/\sigma^2) \quad if \quad i \neq j}_{0 \qquad otherwise}$$

2. Calculate a diagonal matrix D with $D_{ii} = \sum_j (A_{ij})$.

3. L is defined as $L = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$.

4. Find x_1,x_2,. . .,x_k, the k largest eigenvectors of L (chosen to be orthogonal to each other in the case of repeated eigenvalue), and form the matrix $X = [x_1 x_2 ... x_k] \in \Re^{n \times k}$ by stacking the eigenvectors in columns.
5. Form the matrix Y from X by renormalizing each of X's rows to have unit length (i.e. $Y_{ij} = X_{ij}/(\sum_j X_{ij}^2)^{1/2}$)
6. Treating each row of Y as a point in $\Re^K$, cluster them into k clusters via Extened K-means.
7. Finally, assign the original point s_j to cluster j iff row i of the matrix Y was assigned to cluster j.

In this procedure, σ^2 is a parameter controls how rapidly the affinity A_{ij} falls off with the distance between s_i and s_j. However, a fixed parameter may lead to problems when the numbers of documents and variances in the topics are various. For example, even if we have topics A and B, where topic A has larger variance and documents than B, then we get a document d which talks about A and B at the same time and assume that in d topics A and B have the same importance. For human, if creating new cluster is not allowed, we may prefer clustering d into A because it has a larger variance and it seems containing more property. So we introduce an auto-adaptive parameter selection.

To make the parameter σ to be adaptive, we use the following method to select σ parameter for point i and j, which is similar to the work in [7].

The affinity between points s_i and s_j is written as $A_{ij} = exp(-\frac{|s_i - s_j|^2}{\sigma_i \sigma_j})$, where the σ_i and σ_j are in a local scale $\sigma_i = \sqrt{\frac{1}{n}\sum_n |s_i - s_j|^2}$. This method allows σ to be automatically adjusted according to local variance, that is, no parameter setting is required.

As mentioned above, we use an *Extended K-Means* which allows to make new clusters so that we can obtain a good result with smaller distortion. The algorithm is as follows:
1. Given an integer K as an initial cluster number, K centers are randomly selected.
2. We calculate the squared errors with $J = \sum_{C_k \in CLUSTERS} \sum_{\forall x_i \in C_k} (|c_k - x_i|^2)$, where c_k is the center of cluster C_k, and x_i is a point in C_k.
3. For each point, a cluster with the smallest distance to the point is identified.
4. The variance of the cluster with this point added is calculated. If the variance is larger than a threshold, a (sigleton) cluster with the point is newly created. If otherwise, the point is merged into to the cluster.
5. All empty clusters are deleted. Then, for each remaining C_k, its center vector c_k is re-calculated by $c_k = \frac{1}{n}\sum_{x_i \in C_k} x_i$, where n is the number of points in C_k.
6. If the change of squared error is smaller than a given threshold, the clusters are output and the process is terminated. If otherwise, return to 2.

4. Associative Search by Mediators

Let $\mathcal{P} = (P, V, R_P)$ and $\mathcal{D} = (D, V, R_D)$ be a person-term relation and a document-term relaion, respectively, where P is ranked in descending order of their mediator levels. Since we are concerned with concepts in different formal contexts, the derivation operator in a context $\mathcal{C}$ is often explicitly denoted by $'^{(\mathcal{C})}$ or $''^{(\mathcal{C})}$.

Our associative search is perfomed as follows.

Identifying Query Concepts:

A query Q a user interested in is given as a set of keywords (terms) $K \subseteq V$. We can consider a *query concept* in each context. They are defined as $C_Q^{\mathcal{P}} = (K'^{(\mathcal{P})}, K''^{(\mathcal{P})})$ in $\mathcal{P}$ and $C_Q^{\mathcal{D}} = (K'^{(\mathcal{D})}, K''^{(\mathcal{D})})$ in $\mathcal{D}$, respectively.

Getting User Interest:

In order to get a user's interest, we interactively ask the user to express his/her preference on the persons related to the query, that is, the persons in $K'^{(\mathcal{P})}$. Particularly, the user is asked his/her preference only on the Top-N ranked persons with higher mediator levels.

Let $M \subseteq P$ be the list of Top-N ranked mediators in P. For the query concept in $\mathcal{P}$, $Q_{\mathcal{P}} = (P_Q, T_Q)$, according to the user interest, for each person $p \in (P_Q \cap M)$, the user assigns a sign $+$, $-$ or $*$ to p for "*favorite*", "*dislike*", and "*don't care*", respectively, where the sign given to p is referred to as $sign(p)$. Assuming $*$ as default sign to each $p \in P_Q \setminus M$, we can devide P_Q into three groups, $POS = \{p \in P_Q \mid sing(p) = +\}$, $NEG = \{p \in P_Q \mid sing(p) = -\}$ and $DC = \{p \in P_Q \mid sing(p) = *\}$.

Finding Concepts Consistent with User Interest:

Let $Q_{\mathcal{P}} = (P_Q, T_Q)$ be a query concept in $\mathcal{P}$, where P_Q is devided into POS, NEG and DC based on the user interest. Then, a concept $B = (P_B, T_B)$ in $\mathcal{P}$ is said to be *consistent* with the user interest if and only if $P_B \subseteq P_Q$, $P_B \supseteq POS$, and $P_B \cap NEG = \emptyset$.

Since there are in general several consistent concepts, we try to extract maximally general ones among them. That is, our task is to find every maximal concept C_B in $\mathcal{P}$ which is a sub-concept of the query concept $Q_{\mathcal{P}} = (P_Q, T_Q)$ and whose extent must subsume POS and exclude NEG.

In order to find C_B satisfying the constraints, we can basically expand POS by adding a person in DC step by step in a depth-first manner. More precisely, for the closure of person set X_i such that $(POS \cup DC) \supseteq X_i \supseteq POS$, we expand X_i by adding a person $xin(DC \setminus X_i)$ and compute the closure $X_{i+1} = (X \cup \{x\})''^{(\mathcal{P})}$. Then we check whether $X_{i+1} \subseteq (POS \cup DC)$ or not. If yes, X_{i+1} is tried to further expand by adding a person in $DC \setminus X_{i+1}$. If otherwise, X_{i+1} is discarded and X_i is expanded with another person by backtrack because X_{i+1} includes some person in NEG or $(X_{i+1}'^{(\mathcal{P})}, X_{i+1})$ is not a sub-concept of $Q_{\mathcal{P}}$. Such an expansion process recursively iterated untill no closure remains to be expanded.

Identifying User Aspects:

If we can observe a concept which corresponds to the user aspect and bridges Q and C in some sense, it would be reasonable to accept similarity between Q and C under the aspect. In order to realize this kind of associative search, we here formalize a user aspect as a concept reflecting the user interest.

Let $B = (P_B, T_B)$ be a (maximal) concept in $\mathcal{P}$ which is consistent with the user interest. Since B is a sub-concept of $Q_{\mathcal{P}} = (P_Q, T_Q)$, $T_Q \subseteq T_B$ holds. Therefore, $A = T_B \setminus T_Q$ can be viewed as the set of attributes (terms) which can implicitly characterize the user interest.

In order to find similarity of concepts in $\mathcal{D}$ under the aspect reflecting the user interest in persons, we consider a concept in $\mathcal{D}$ which is defined based on A. Formally speaking, if A is a closure in $\mathcal{D}$, that is, $(A'^{(\mathcal{D})}, A''^{(\mathcal{D})} = A)$ is a concept in $\mathcal{D}$, then we regard the concept $C_A = (A'^{(\mathcal{D})}, A)$ as an aspect reflecting the user interest in persons.

Extracting Conditionally Similar Concepts w.r.t. User Aspect:

We first define a similarity between concepts without any conditioning. Our similarity can be defined based on the notion of *bond* [4], an extension of *Jaccard Coefficient.*

Let $C_1 = (X_1, Y_1)$ and $C_2 = (X_2, Y_2)$ be a pair of concepts. Then a similarity between C_1 and C_2, denoted by $sim(C_1, C_2)$, is defined as $sim(C_1, C_2) = \frac{|X_1 \cap X_2|}{|X_1 \cup X_2|}$.

The measure is extended for *conditional similarity* between concepts. Let $C_1 = (X_1, Y_1)$ and $C_2 = (X_2, Y_2)$ be a pair of concepts. For a set of attributes R, a similairty between C_1 and C_2 with respect to R, denoted by $sim(C_1, C_2|R)$, is defined as $sim(C_1, C_2|R) = \frac{|\bigcap_{y \in Y_1 \cup Y_2}(R' \cap y')|}{|\bigcup_{y \in Y_1 \cup Y_2}(R' \cap y')|}$.

Let δ be a given threshold for the minimum similarity value. For a pair of concepts C_1 and C_2 and a set of attributes R, if $sim(C_1, C_2|R) \geq \delta$, C_1 and C_2 are said to be *conditionally δ-similar* with respect to R.

For a query concept Q and the user aspect C_A, if C_A bridges Q and a concept C in some sense, then it would be natural to find similarity between Q and C under the aspect. We call such a C_A a *bridge concept* for Q and C. It is formally defined as follows.

Definition 1 (δ-Bridge Concept)
Let $C_R = (X_R, Y_R)$ and $C_L = (X_L, Y_L)$ be a pair of concepts. For a threshold δ, if a concept $C_W = (X_W, Y_W)$ satisfies the following conditions, then C_W is called a *δ-bridge concept* between C_R and C_L w.r.t. Y_W.
Structural Constraint: $X_R \cap X_W \neq \emptyset$ and $X_L \cap X_W \neq \emptyset$.
Conditional Similarity Constraint: C_R and C_L are conditionally δ-similar with respect to Y_W. ■

Regarding the user aspect $C_A = (A'^{(\mathcal{D})}, A)$ as a bridge concept, for the query concept $Q_{\mathcal{D}} = (K'^{(\mathcal{D})}, K''^{(\mathcal{D})})$ in $\mathcal{D}$, our target we try to find is a concept $C_{target} = (T''^{(\mathcal{D})}, T)$ such that (1) $T \cap A = \emptyset$, (2) $K''^{(\mathcal{D})} \cap T = \emptyset$, (3) $K'^{(\mathcal{D})} \cap T'^{(\mathcal{D})} \neq \emptyset$ and (4) $bond(K''^{(\mathcal{D})} \cup T|A) \geq \delta$. Particularly, we try to extract maximal ones satisfying the constraints. It should be noted here that the constraints (1) and (2) are not included in our original definition of δ-bridge concepts. They are assumed to obtain more interesting associations of terms. Moreover, from a computational view point, they can restrict our search space for efficient computation.

In order to obtain our target, we recursively expand a closure of terms in depth-first manner. Let $X_i \subseteq V$ be the closure of a set of terms such that $X_i \cap (K''^{(\mathcal{D})} \cup A) = \emptyset$. For a term $x \in V \setminus (K''^{(\mathcal{D})} \cup A \cup X_i)$, we check whether $X_{i+1} = (X_i \cup \{x\})''^{(\mathcal{D})}$ as T satisfies all of the four constraints. If X_{i+1} does not satisfy (1) or (2), X_{i+1} can be discarded. If X_{i+1} does not satisfy (3), then any expansion of X_{i+1} also violates the constraint. Therefore we can stop expanding X_{i+1}. If the constraint (4) cannot be satisfied for X_{i+1}, any expansion of X_{i+1} can be pruned. This is because the similarity measure is monotonically decreasing as a set of attributes (terms) becomes larger. If X_{i+1} is discarded, X_i is tried to expand with another term in $V \setminus (K''^{(\mathcal{D})} \cup A \cup X_i)$ by backtrack.

On the other hand, all of the constraints are satisfied for X_{i+1}, then $(X'^{(\mathcal{D})}_{i+1}, X_{i+1})$ becomes a candidate of our target. Then X_{i+1} is further tried to expand with a term in $V \setminus (K''^{(\mathcal{D})} \cup A \cup X_{i+1})$. Such an expansion process is recursively iterated until no closure ramains to be examined.

5. Summary

We have presented our framework of associative search. A remarkable point is that our association is controlled by user interests in good mediators. We have proposed a method for ranking persons according to their mediator levels. Particularly, their mediator levels are defined based on clusters of terms obtained by an extended K-means algorithm. We have designed a computational procedure for our association search and designed depth-first algorithms for extracting our target concepts.

References

[1] A. Takano, Y. Niwa, S. Nishioka, M. Iwayama, T. Hisamitsu, O. Imaichi and H. Sakurai, Information Access Based on Associative Calculation, Proc. of SOFSEM'00, 187 - 201, 2000.

[2] Y. Okubo and M. Haraguchi, An Algorithm for Finding Indicative Concepts Connecting Larger Concepts Based on Structural Constraints, Contributions to ICFCA 2011, pp. 53 – 68, 2011.

[3] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Springer, 1981.

[4] E. R. Omiecinski, Alternative Interest Measures for Mining Associations in Databases, IEEE Trans. on KDE, 15(1), 57 – 69, 2003.

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

[6] A. Ng, M. Jordan, Y. Weiss, On Spectral Clustering: Analysis and an algorithm, Proc. of NIPS'01, 849 – 856, 2001.

[7] L. Zelnik-Manor and P. Perona, Self-Tuning Spectral Clustering, Proc. of NIPS'04, 1601 – 1608, 2004.