Information Retrieval Based on Formal Concept Analysis

ZHI Dongjie
School of Compute Science and Technology, Henan Polytechnic University, Jiaozuo, Henan, P.R.China, 454000
zhihuilai@hpu.edu.cn

Abstract: In information retrieval, in order to improve the recall and precision, similar concepts are often processed during the information retrieval. With this regard, the possibility of evaluating concept similarity is acquiring an increasing relevance, since it allows the identification of different concepts that are semantically close. For every concept consist of two parts which include objects and attributes, so the computing model also has two parts. First part deals with the objects similarity by using hierarchy structure of concept lattice. The second part deals with the attributes similarity by using the expert’s knowledge. Beside this, the depth also has relevance with concepts similarity. So the result is altered with concepts’ depth. Experiment shows using concept lattice could improve information retrieval’s recall and precision.

Keywords: Concept lattice, information retrieval, concept similarity , generation operation, expansion operation

1 Introduction

Formal concept analysis is a research area of knowledge representation, and its core data structure is concept lattice, which is a description of the relationship between the concepts and to a certain extent it is a highly simplified description of the objective world. The biggest advantage of this simplification brings about good mathematical properties. R. Wille [1] based on this simplified system and studies the concept of the orderly nature of the lattice algebra and formal concept lattice is isomorphic to the nature of the background, and founds a field of formal concept analysis. It is due to this simplification, making the concept lattice can not simply be understood as part of the objective world model, but more as an artificial data set derived from a number of representations.

Because of its conceptual structure, which is facilitating the development and discussion, in a sense, the concept lattice has become a means for external recognition [2]. Dau [3] studied the formal concept analysis based on graph representation for reasoning. Wille [4] summarizes the formal concept analysis for knowledge management in the ten features. Stumme [5] pointed out the formal concept analysis of the five principles of knowledge representation: as a medium of human grammatical knowledge representation, as a set of appointments scheduled of ontology, as agent, as intelligent reasoning of the local theory and as a practical and effective calculation medium. Kollewe et al [6] developed Toscana system also confirms the concept of formal concept analysis is fit for knowledge discovery. This article will use the formal concept analysis to establish information retrieval model to improve information retrieval precision and recall, the core technology is the similarity calculation of concepts.

2 Introduction of formal concept analysis

Formal concept analysis, which is proposed by R. Wille [1], is founded on a basis of order theory and lattice theory, and is a mathematical structure which depicts relationship between objects and attributes according to basic information provided by data base.

Definition 1 Given a formal context $K=(U,A,I)$, $X \subseteq U$, $B \subseteq A$, if $X$ and $B$ satisfy $f(X)=B$, $g(B)=X$, and $f(A):=\{m \in M| \forall g \in A, gIm\}$, $g(B):=\{g \in G| \forall m \in B, gIm\}$, we call the pair $(X,B)$ is a concept of formal context $K$. $X$ is called extent of concept $(X,B)$, $B$ is called intent of concept $(X,B)$. $L(U,A,I)$ is the set which contains of the concepts of $K$, $L(U,A,I)=\{(X \times B) \in (U \times A); f(X)=B, g(B)=X\}[1]$. 

1
Definition 2 Given a formal context $K = (U, A, I)$, $(X_1, B_1), (X_2, B_2) \in L(K)$, if $X_1 \subseteq X_2$ or $B_2 \subseteq B_1$, then call $(X_1, B_1)$ is a son concept of $(X_2, B_2)$, and denote as $(X_1, B_1) \leq (X_2, B_2)$. Apparently, $L(K)$ forms a lattice, which is called concept lattice\(^{[1]}\).

Given two concepts $(X_1, B_1), (X_2, B_2)$ of a formal context $K = (U, A, I)$, we have: If $X_1 \subseteq X_2$, then $f(X_2) \subseteq f(X_1)$, for $X_1, X_2 \subseteq U$; If $B_1 \subseteq B_2$, then $g(X_2) \subseteq g(X_1)$, for $B_1, B_2 \subseteq A$.

From the definition and nature of the concept lattice, we can see, concept lattice not only give the precise definition of the concept, but also more importantly, describes the inheritance relationship between concepts, which provides a good data structure for concept similarity measure.

3 Concept similarity calculations

3.1 Basic idea of concept similarity calculations

The concept of similarity can be measured by the distance of the concepts hierarchy in the concept lattice. Intuitively, the larger the distance is, the lower their similarity; the other hand, the smaller the distance the two concepts, the greater the degree of their similarity. Distance is 0, the similarity is 1; concept of distance is infinity, the similarity is 0; similarity is monotonically decreasing function of the concept of distance.

3.2 Concept similarity calculations

The distance between concepts can be measured by objects and attributes in the concept lattice: the greater the distance of the two concepts, the less same objects and same attributes shared by the two concepts; the other hand, the smaller the distance of two concepts, the more same objects and same attributes shared by the two concepts.

In \(^{[7]}\), the calculation of attribute similarity, not only calculate similarity of same attributes, but take different attributes into account, and this is not appropriate in a uniform background. Attribute is used to distinguish different objects, if attributes are similar, then these two attributes don’t have the ability to differentiate objects, therefore, there can be no two attributes are the semantic cross. So in the calculation, just take same attributes into account is enough.

Definition 3 In a given domain ontology and several formal contexts $(U_i, A_i, I_i) \ i=1\ldots k$, the similarity of the concepts $(X_1, B_1), (X_2, B_2)$ is:

$$
Sim((X_1, B_1), (X_2, B_2)) = \left( \frac{|X_1 \cap X_2|}{m} * a + \frac{|B_1 \cap B_2|}{n} * b \right) * (1 + c)^{\left(1 - \frac{m}{n}\right)},
$$

$a + b = 1, c > 0, m = \max (|X_1|, |X_2|), n = \max (|B_1|, |B_2|)$.

According to the dual principles of concept lattices \(^{[1]}\), objects and attributes have the same status of a concept, then we have $a = b = 0.5; l_1, l_2$ are the depths of $(X_1, B_1), (X_2, B_2)$ in the concept lattice; $c$ is the amendment which reflects the depth of the impact of similarity, and we take $c = 0.01$.

3.3 An example

The example’s formal context is from literature \(^{[7]}\), which is shown in Table 1. It describes some famous European cities: Athens(A), Courmayeur(C), Innsbruck(I), London(L), Paris(P), Reyjkjavik(Re), Rome(Ro), and they ahhs attributes: Archeological-site(Arc), Beach(Bea), Capital(Cap), Euro(Eur), River(Riv), Skiing-area(Ski). The concept lattice of this formal context is in fig.1.

According to Definition 3, we can calculate several concepts similarity as following:

$$
\begin{align*}
\text{Sim}(&\{L, P, Ro\}, \{\text{Cap, Riv}\}) = 0.61; \\
\text{Sim}(&\{L, P, Ro\}, \{\text{Cap, Riv}\}) = 0.70; \\
\text{Sim}(&\{P, Ro\}, \{\text{Cap, Eur, Riv}\}) = 0.38; \\
\text{Sim}(&\{Ro\}, \{\text{Arc, Bea, Cap, Eur, Riv, Ski}\}) = 0.21.
\end{align*}
$$

It can be seen from the results, sibling concepts’ similarity is smaller than father-son concepts, and
similarity of the concept of grandparents and grandchildren is smaller than father-son concepts, and similarity of the concepts which have no direct genetic relationship is smallest. Several formal contexts can be combined to form a single formal context, then we generate concept lattice and carry on the calculation of concept similarity.

<table>
<thead>
<tr>
<th>Table 1 Formal context of “European cities”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>I</td>
</tr>
<tr>
<td>L</td>
</tr>
<tr>
<td>P</td>
</tr>
<tr>
<td>Re</td>
</tr>
<tr>
<td>Ro</td>
</tr>
</tbody>
</table>

Fig.1 concept lattice of formal context “European city”

4 Information retrieval based on concept lattice

In information retrieval, in order to improve recall, if retrieve none relevant information by given keywords, then needs to carry out generalization or expansion operation of keywords[8], and go on retrieve information by similar concepts.

Retrieval algorithm:

Retrieve (Key)
{
    Find relative data and save to Result;
    If Result! =Null
Then return Result;
Else
{New_Key=Find_Parent_Concept (Key);
  If Sim (New_Key, Key)> ε //similarity calculation
  Then
  {Key=New_Key;
   Retrieve (Key) ;} //recursive call
  } //generation operation
{New_Key=Find_Sibling_Concept (Key);
  If Sim (New_Key, Key)> ε //similarity calculation
  Then
  {Key=New_Key;
   Retrieve (Key) ;} //recursive call
  } }//expansion operation

Combined with the author's research area, establish domain ontology "ontology", including: body, event, theme, event-based ontology, domain ontology, and some other relevant concepts of the background. We develop a literature search tools combined with the above algorithm, and experiments on the 2463 literatures which download from "National Knowledge Infrastructure" (CNKI). Experiment shows recall is improved. For example: when not using the above algorithm, enter "event ontology", there is no literature hit back, after adopts the above algorithm, 19 articles hit back, only 3 of them is not relevant.

5 Conclusion

In order to improve the recall and precision, similar concepts are often processed during the information retrieval. In this paper, we use Formal Concept Analysis's core data structure, namely concept lattice to aid the information analysis process. For every concept consist of two parts - objects and attributes, so the computing model should have two parts. First part deals with the objects similarity by using hierarchy structure of concept lattice. The second part deals with the attributes similarity by using the expert's knowledge. Beside this, the depth also has relevance with concepts similarity. In information retrieval process, we use generation operation and expansion operation about the given key words to improve recall and precision. Experimental results show that the proposed model fully takes into account the impact of distance on the concept similarity calculation, as well as hierarchy structure of concepts; the results can meet the requirements of practical applications.

References

[5]. Stumme, G. Formal concepts analysis on its way from mathematics to computer science [A]. In U. priss, D. Corbett, & G. Angelora (Eds.), Conceptual structures: Integration and Interfaces, 10th

