Wu, Y. (2000). Automatic concept hierarchies development: A revised subsumption approach. . 11th ASIS SIG/CR Classification Research Workshop, 137-140. doi:10.7152/acro.v11i1.12776

Automatic Concept Hierarchies Development: A Revised Subsumption Approach

Yi-Fang Wu
Ph.D. Student, School of Information Science & Policy,
State University of New York at Albany, Albany, NY 12222

Abstract

In this study, the original subsumption rule proposed by Sanderson and Croft is revised. Different thresholds are used to observe how shapes of a concept hierarchy change. Ranking among children concepts is available based on sorted subsumption data. This study also explores three potential usages of concept hierarchies. 1. As an overview to a document collection. 2. As a tool to compare different document collections in the same domain. 3. As a tool to observe evolution trends in a domain.

1. Introduction

More and more books, documents, materials are now available to everyone on the Internet. Information seekers must be provided with tools that help them to have a better understanding of material retrieved. This tool should provide useful information such as an overview of documents retrieved, concept relationships, and ranking of concepts to users. Current clustering and classification techniques have many problems, such as ambiguous terms are only listed in one place on the hierarchical clustering trees, or pre-defined classification structures like Library of Congress Subject Headings do not keep up with current development of a field.

The purpose of this study is to develop a feasible and reliable technique for creating a hierarchical representation of concepts derived directly from documents. Subsumption rule was found feasible to accomplish this task (Sanderson & Croft, 1999). A concept hierarchy derived directly from text is better than using pre-defined concept hierarchies, for it exactly reflects the content and how concepts are used in the text. It also resolves the ambiguous terms problem, since such terms will be listed in multiple places on the concept hierarchy. However, flaws were found in the original rule. In this study, flaws of subsumption rule are revised and potential usages of concept hierarchies are explored.

2. Original and Revised Subsumption Rule

2.1 The original rule

The definition of original subsumption rule is simple. X and Y are two terms. If Y appears in a subset of documents which X appears in, then X is said to subsume Y. The rule can be represented as follows: $\mathbf{P(x|y)=1, P(y|x)<1}$ —— **(1)**

ISSN: 2324-9773

Wu, Y. (2000). Automatic concept hierarchies development: A revised subsumption approach. . 11th ASIS SIG/CR Classification Research Workshop, 137-140. doi:10.7152/acro.v11i1.12776

Because X subsumes Y and also because X' s number of occurrence is higher, in the hierarchy, X is the parent of Y. Sanderson and Croft noticed that many term pairs were failing to be included because a few occurrences of Y did not co-occur with X. So, the subsumption rule was redefined as follows: **$P(x|y)>=0.8, P(y|x)<1$ ——— (2)**

The cut-off value 0.8 was chosen through informal analysis of subsumption term pairs. The flaw was found here. It is very possible that the new rule will include $P(X|Y) = 0.8$, $P(Y|X) = 0.8$. In this case, X does not subsume Y. Moreover, if $P(X|Y) = 0.8$ and $P(Y|X) = 0.9$, Y subsumes X. Therefore, Sanderson and Croft' s subsumption rule is not valid in every situation.

2.2 The Revised Rule

To ensure X subsumes Y, subsumption rule needs to be modified as follows:

$P(X|Y) >= N > P(Y|X), 0 < N < 1,$ ———(3)

The cut-off point value is represented as N in the revised rule. A greater cut-off value result s in a smaller, rigid hierarchy, for there will be fewer term pairs fulfill the threshold. On the contrary, the smaller cut-off value results in a bigger hierarchy. It is suggested that the cut-off point should be determined by the user, based on the purposes of his/her task.

3. Research Possibilities

3.1 The effect of cut-off points on the resulting concept hierarchies

The cut-off point affects the shape and number of concepts on resulting concept hierarchy. One of the important aspects of this study is to determine if a general cut-off point can be found.

3.2 Ranking

One by-product of subsumption test is the ranking of concepts. Suppose parent concept X has 10 children concepts. By ranking P (child | Parent X) value, it is very easy to determine which child concept is the most important one among children concepts of parent concept X. If concepts are arranged systematically on the hierarchy, based on ranked P (X | Y) value, it is feasible to determine which part of the concept hierarchy consists of major concepts in the document set.

3.3 Exploring Potential Usages of Concept Hierarchies

3.3.1 As an overview to a document collection.

A concept hierarchy can provide an information seeker with a "rough idea" about the document set. In other words, a concept hierarchy provides an overview to documents. (Sanderson & Croft, 1999) The following should be examined: term coverage, and the accuracy of term relations on the hierarchy.

3.3.2 As a tool to compare different document collections in the same domain.

ISSN: 2324-9773

Wu, Y. (2000). Automatic concept hierarchies development: A revised subsumption approach. . 11th ASIS SIG/CR Classification Research Workshop, 137-140. doi:10.7152/acro.v11i1.12776

A concept hierarchy can be used to compare documents. For example, comparing two books published by two different authors that both have the title "information retrieval". By creating one concept hierarchy for each, the concept hierarchy can be used to compare how two authors introduce the field of information retrieval.

3.3.3 As a tool to observe evolution trends in a domain.

Back the above example, if two "information retrieval" books were written 10 years apart, then concept hierarchies can also be used to observe the evolution of a field. The following questions should be examined on the resulting concept hierarchies: What are new terms? What terms disappeared? Do the changes reflect the evolution trend of the domain? If a term appears in both document sets, has its term relations with other terms in the concept hierarchy changed?

4. Conclusion

This study focuses on applying revised subsumption rule to create a better classification method for knowledge organization. Three potential usage studies help concept hierarchy to be better used for concept classification for document sets.

Reference:

Sanderson, M., & Croft, B. (1999). Deriving Concept Hierarchies Form Text. *Proceedings on the 22nd annual international ACM SIGIR conference on Research and development in information retrieval , 1999,* , 206 - 213.

ISSN: 2324-9773

Wu, Y. (2000). Automatic concept hierarchies development: A revised subsumption approach. . 11th ASIS SIG/CR Classification Research Workshop, 137-140. doi:10.7152/acro.v11i1.12776

Questions:

1. Is the revised subsumption rule valid?
2. What factors affect the accuracy of a concept hierarchy generated by using the revised subsumption rule? (such as choices of term vs term phrases)
3. Can concept hierarchies be used other than potential usages I mentioned?

ISSN: 2324-9773