Adaptive Semantic Measurement for Information Filtering

(Work-in-Progress)

Glenn Boardman
Department of Computer Science and Computer Engineering
La Trobe University
Bundoora, Melbourne
VIC 3086, AUSTRALIA
gcboardman@students.latrobe.edu.au

Hongen Lu
Department of Computer Science and Computer Engineering
La Trobe University
Bundoora, Melbourne
VIC 3086, AUSTRALIA
helu@css.latrobe.edu.au

ABSTRACT

With the volume of information on the Internet growing at an exponential rate, the needs of users to have their search results effectively filtered is increasingly important. This paper examines how a tree threshold function can be used in an information filtering agent (IFA) to extend the original keyword search to cover other related words within the domain, creating a keyword weighted semantic tree. The examination in this paper also considers how the metrics of the tree structure (shape, size, weights) influence the choice of related words for use in the extended search and what advantage this has over traditional methods. Further, that using a reduced word tree, which has been pruned using the tree pruning algorithm produces a significant increase in the number of profitable results for the user. Using these factors the analysis demonstrates equal accuracy to the benchmark comparison IFA but with increased efficiency.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering

Keywords

Ontology, Agents, Semantic Measurement

1. INTRODUCTION

As information on the Internet continues to grow at an exponential rate, the ability to search web pages and return meaningful results becomes a more daunting task. This problem arises because current search engines do not take the keyword domain into account. To alleviate this problem, one approach is to use an Information Filtering Agent (IFA) to automate the process of filtering results . An IFA not only analyzes the number of occurrences and locations of keywords within a document, but also analyzes the relationships between keywords. These relationships include the study and detection of related words from a search and how they may be used to extend the specified domain and facilitate a more meaningful search. In this paper, we propose an adaptive function to measure the semantic distance of keywords taking into account of domain knowledge and the shapes of semantic relations (ontologies). This function serves as an important role to improve the accuracy of information filtering. Experiments and analysis show promising results using this function in IFA.

Conference name: Infoscale 2007, June 6-8, 2007, Suzhou, China Copyright number (LaTeX): 978-1-59593-757-5

2. THRESHOLD FUNCTION FACTORS

The problem with most current implementations is the lack of domain knowledge and inadequate word tree scoring functions. Ontology scores are usually either static or are specific to only a few words derived from a semantic tree. Weighted semantic word trees are an excellent option as they can cater for topic variations, however the tree structures are usually manually designed and the ontology scores are determined and influenced by the programmer's perception. The primary focus of this research is to design and develop an adaptive semantic measurement function using geometric analysis (eg shapes, dimensions, scope). A word tree threshold function is needed to analyse the formed tree shapes and other attributes when a keyword is searched on the Internet and to determine the level of derived focus nodes for use when refining the search. The tree threshold function will use an existing semantic network such as WordNet [1]. This section covers the threshold generation function and the factors that have been taken into consideration, how they effect the overall operation of the IFA and whether they are profitable in refining search results. The way a parameter is used is just as important as its raw result. The parameters in this function are used in two manors: as a raw statistic (a number or set of numbers) or as a scaler (a multiplier used to scale a factor, increasing or decreasing its significance)

2.1 Average Tree Weight Per Level

The issue with many possible factors is that their characteristics may be replicated by any number of different focus words. The advantage of the average weight per level is that it is specific to a weighted tree. Therefore this factor is individualized to a particular keyword tree.

2.2 Tree Shapes

The parameter of the tree shape is the most direct and obvious of the factors in use. This could also be described as the characteristic decay of a tree shape as it predicts the likely score decay the tree will undertake as it grows.

2.3 Average Tree Width and Height

This metric holds two functions. Firstly, it is another characterizing parameter which is specific to the ontology tree which has been created. And secondly this tree metric gives scope to the scale and size of the tree created. A larger tree with more nodes can afford to have more nodes pruned from it.

2.4 Individual Branch Heights/Depth

In a similar manor to the average tree width, this metric gives scope to the size of the ontology word tree. By looking at a particular branch, you can observe how far it extends and therefore how abstract or specific words become as they approach the base of the tree.

2.5 User Input - Pruning

User input can be an important factor in pruning the ontology trees. Every user may not have the same requirements for filtering as another user, thus some form of input is generally useful in restricting the search domain. By using this factor, results can be easily catered to a particular search need.

2.6 Synonym Set (Synset) Size

Synonyms, being the closest relation to the specific keyword provide the best possible extended words and indication of how many tree levels could be included in the extended search tree.

3. SURMISED THRESHOLD FUNCTION

To achieve the optimal output from the threshold generator, it is important to incorporate the discussed factors in an appropriate way. The base of the threshold generation function centers around two factors being the average weight per level and the tree shapes. The average weight is the most important factor as it outlines the weighting of the word tree, but the tree shape is also important because it indicates the individuality of the topic. The formal definition of this function is as follows: $T_s = U_i \times \frac{\sum_{i=0}^{I-1} BH_i}{A_h} \times \vdash_{height=0}^{N-1} (\frac{AW_{height} + TS_{height}}{2} \times \frac{\sum_{n=0}^{N-1} \frac{SS_{height}}{length(SS_{height})}}{length(SS)})$

The core of this function is based around the "a verage branch height" and the "tree shape decay" factors which are taken as a 50/50 value. As discussed above, these two factors are used as base factors since they provide the most meaningful description of the tree shape. Given any tree height n, these two base factors are evaluated and then multiplied with the other factors to vary the thresholds generated.

4. EVALUATION AND ANALYSIS

Each test was run on the SQL data set in order to generate a results set. From this data a table has been populated and shows the generated page rank, page index and percentage of relevant documents returned by the page scoring agent (in contrast to the comparison function and human ranking). This testing scheme ensures that if there is a detectable difference in the threshold functions results, it can be properly identified.

4.1 Hyponym Evaluation

The first part of evaluating the hyponym function was to find the most beneficial balance between factors. Analysis of server combinations proved that a combination of 60% average tree weight and 40% characteristic decay yielded the most pleasing results. Testing was conducted using six methods. Searching and scoring based on a single keyword, searching based on a full semantic tree with each node having unitary weight, searching based on a full semantic tree with each node having full weight, searching based on a function pruned semantic tree with each node having unitary weight, searching based on a function pruned semantic tree with each node having full weight. Evaluations of these results showed that using the fully weighted semantic tree returned the greatest number of relevant results at 95% relevance in the top 15 results. As predicted, the worst performance came from the single word search and full tree, unitary weight at 85% relevance score. The fully weighted, function pruned tree yielded 92% relevance with a significant decrease in the required processing time.

4.2 Hypernym Evaluation

The level of word abstraction plays an important part in the threshold generation function. At the top of all hypernym word trees are terms such as entity or object, which will not usually contribute profitably to a users search. Thus the aim of this function is to gauge how rapidly terms in the hypernym tree become too abstract by examining several different words and their structure. It was concluded that one third of the mean height of the tree was sufficient to include most profitable words. However as stated earlier, a very specific search word may have many useful hypernyms, so one third of them would not be enough. Consequently the function is catered towards low and moderate height hypernym trees whose root word is not very focused towards a particular domain.

$$height = ceil[\frac{avg}{floor(\frac{avg}{6} + 2)}] \quad (1)$$

To achieve this a linear expression was chosen, as below, using the mean height of all branches as the starting height, then divided by the linear decay.

$$avg = \frac{\sum_{n}^{i=0} branchheight_i}{no.branches} \quad (2)$$

This value was adopted after testing various linear expressions, and their success when applied to several different hypernym word structures. This formula allows for the inclusion of more nodes when analysing larger trees, but has a greater impact on small structures with a fast decay to abstract nodes. As hypernyms rarely exists within a document without associated hyponyms, the analysis of the hypernym function has been incorporated in the hyponym analysis section.

5. CONCLUSION

This paper demonstrates that an adaptive semantic function can improve searching and filtering of web pages. Evaluation results show a marked improvement using extended word searching as opposed to single word processing. The IFA performed as well as the comparison/benchmark IFA, due to the reduction of unprofitable related words, there was a marked increase in the accuracy and profitability of the resulting webpages in the chosen domain. The analysis demonstrates that the threshold generation function's maximum performance was achieved when using a tree with an even distribution of nodes, its application on other tree shapes still provided positive results with only a slight decrease in relevance and processing efficiency.

6. REFERENCES

[1] G.A. Miller, Beckwith R., C. Fellbaum, D. Gross, and K.J. Miller. Introduction to wordnet: An on-line lexical database. *Journal of Lexicography*, 3(4):234–244, 1990.