An empirical study of approaches to determining the semantic relatedness of web pages

Shantha Jayalal Pearl Brereton Chris Hawksley
School of Computing and Mathematics
Keele University, Keele, Staffordshire
ST5 5BG, United Kingdom
{*shantha, pearl, chris*}@cs.keele.ac.uk

Abstract

With increasing numbers of Web users, there is a necessity to improve their Web site navigation experience over the Internet and a range of Web applications have emerged recently for this purpose. Many researchers have stressed the importance of identifying semantic relatedness of Web pages in such Web applications as Web site navigation, automatic tour generation and adaptive Web applications.

One approach to identifying semantic relatedness between documents is to use lexical databases and lexical chains. For example, an approach using lexical chains has been proposed by Green for identifying paragraph similarity in a document [1]. However, due to the unacceptable length of time needed for lexical chaining and the difficulty of global representation of documents, Green used *synset weight vectors* to compare semantic relatedness between two documents. But his approach to identifying paragraph similarities can be extended to identify semantic similarities between documents.

In this study, an approach to identifying semantic similarity between Web pages incorporating weighted lexical chains (SR_{WLC}) and document properties based on reiteration, density, length and semantic distance is proposed. The two approaches (the proposed approach and Green's approach - SR_{Green}) were empirically compared by determining the semantic relatedness of Web pages using human subjects.

The research hypothesis of this research is that the proposed approach identifies significantly more semantically related pages with a higher precision than the approach that has been proposed by Green. The null hypothesis is that there is no significant difference in identification of semantically related pages between the two approaches. In this context precision is defined as the proportion of retrieved pages that are relevant.

Web pages belonging to the Department of Computer Science, Keele University are used for the empirical evaluation of the two methods. The semantic relatedness of all pages was identified using both approaches and a Web-based page categorisation exercise using human subjects was carried out for this empirical evaluation. The evaluation is Web based, and therefore can be carried out on the subjects' preferred Web browser at his or her preferred time & place. Therefore the distortion effects are minimised and the results of the evaluation are realistic and also can be reliably generalised to some extent.

An invitation e-mail was sent to twenty subjects during the first week of October 2004 giving a link and guidelines for them to start the experiment. Once the link on the email was clicked, subjects were shown the initial Web page, giving an introduction to the experiment and instructions on how to continue. Twelve out of twenty invited subjects completed the experiment. Two subjects attempted the experiments but couldn't finish because of network problems; another three subjects couldn't finish because of time restrictions, and the other three subjects did not respond at all. Therefore responses from only twelve subjects are used for experimental evaluation.

The Wilcoxon signed ranks test returns a p value of 0.004, indicating that the null hypothesis can be rejected, and that there is evidence to suggest that the SR_{WLC} approach identifies significantly more semantically-related pages with higher precision than the SR_{Green} approach.

The SR_{WLC} approach should be evaluated further using different Web site contents and language styles (e.g. American/British English). It would be interesting to use more subjects from different backgrounds to do the evaluation. This would determine whether the results of the evaluation are influenced by the human subject's background, such as their status (student, staff, or other) his familiarity with the pages of the test database, gender differences or the level of English knowledge. Most importantly the approach is believed to be valid for more general browsing environments than a computer science Website and a wider study is desirable.

REFERENCES

- [1] Green, S. J. (1999), "Building Hypertext Links by Computing Semantic Similarity", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 11 Issue No. 5.