Automatic Hypertext Keyphrase Detection

Daniel Kelleher, Saturnino Luz
University of Dublin, Trinity College
Department of Computer Science, O'Reilly Institute, Trinity College, Dublin 2, Ireland
{dkellehe, luzs}@cs.tcd.ie

Abstract

This paper describes initial experiments in applying knowledge derived from hypertext structure to domain-specific automatic keyphrase extraction. It is found that hyperlink information can improve the effectiveness of automatic keyphrase extraction by 50%. However, the primary goal of this project is to apply similar techniques to information retrieval tasks such as web searching. These initial results show promise for the applicability of these techniques to more far-reaching tasks.

1 Introduction

The associative nature of the Web has been under-exploited so far. This paper describes the initial steps taken towards developing a framework that will take advantage of the associative hyperlink structure of the web to improve information retrieval and document classification. In this project, the automatic keyphrase extraction program, KEA, was adapted for use on web corpora and an extra feature was added that takes advantage of any semantic similarity that may exist between linked web documents.

1.1 KEA

KEA [Witten *et al.*, 1999] is an automatic keyphrase extraction algorithm for documents based on a domain-specific machine-learning model [Frank *et al.*, 1999]. It compiles a list of phrases from a training set of documents with annotated keyphrases and generates a naïve Bayes classifier based on two default features of these phrases:

- TFxIDF [Salton, 1988], which is a value based on the ratio of probability of a phrase appearing in the current document and the probability that it appears in any document, and is given by:

$$TF \times IDF\ (P, D) = \frac{\text{No. of occur. of P in D}}{\text{No. of phrases in D}} \times \log\left(\frac{\text{total number of documents}}{\text{number of docs that contain P}}\right)$$

- Distance - the ratio of the number of words before the first appearance of the phrase in the document and the total number of words in the document.

The classifier is then used to extract potential keyphrases from a test set of documents.

1.2 Hypertext

The concept behind hypertext is that text content (or other media, in fact) is connected by associations or 'links' from document to document, forming a directed graph structure. The associations will usually (although not always) be based on some semantic similarity or relevance (of varying strength) between two documents.

2 Method

The link structure of web documents is included in KEA by introducing the "Semantic Ratio" (SR) feature. SR is similar to the TFxIDF feature, in that it is a frequency ratio. However, the SR of a phrase is calculated by dividing the number of occurrences of that phrase in the current document by the number of times it occurs in all documents directly linked to that document (i.e. those that are the targets of hyperlinks in the document).

$$SR(P, D) = \frac{\text{Frequency of P in D}}{\text{Frequency of P in documents linked to D}}$$

The reasoning behind including this feature in KEA is based on the intuition that the content of a web document is frequently semantically related to its neighbours (in the context of a graph structure, in other words, the documents linked to it) and that the subject matter (identified by the keyphrases) of the document is therefore in some way relative to their contents. The inclusion of the SR feature is a first step in testing, and subsequently, modelling this intuition.

A low SR value (< 1) indicates that a potential keyphrase occurs more frequently in the document's neighbours than in the document itself. The higher the SR value, therefore, the more specific the phrase to this document, relative to its immediate surroundings. Note that this is different to the TFxIDF score as only a subset of the documents are used to compute it, namely those documents that form a localised subgraph with paths of length 1 from the original document.

3 Testing

The new version of KEA with the SR feature included was tested on four web corpora taken from the WWW. These corpora were chosen because a sufficient number of docu-

ments in each site contained annotated keyphrases in the form of the Meta Keyword HTML tag and were therefore suitable for empirical tests on the accuracy of automatic keyphrase extraction.

The corpus was then split into a training set and a test set of roughly equal size, with the restriction that no document in the test set should be linked to a document in the training set. The new KEA algorithm (called KEAWeb) was then trained on the training set and then performed automatic keyphrase extraction on the test set, and the average number of correct keyphrases found in each corpus was recorded and is presented below.

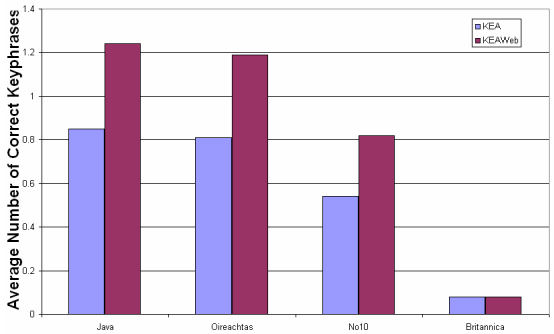

Figure 1: Comparison of KEA and KEA Web

4 Results

Figure 1 shows that the inclusion of the SR feature in the naïve Bayes classifier improves the success of the KEA algorithm by between 45% and 52% in three out of four of the test corpora. KEA performs badly on the Britannica corpus whether the SR feature is included or not. This is due to a number of factors based on the fact that the license-free version of the site severely restricts both the size and the number of links in a document. It can therefore be safely assumed that the corpus is an atypical example of a web site and that the three other corpora are more representative examples.

Initial analysis of the SR distribution in keyphrases suggests that phrases with extremal SR values are more likely to be keyphrases. In other words, phrases that appear frequently in surrounding documents (low SR) have high relevance for the document in question. Also, phrases that occur very rarely in surrounding documents (high SR) will also have high relevance, suggesting that they indicate a topic that is specific to the current document.

5 Conclusions & Future Work

The SR feature shows initial promise as an indicator of semantic relations between linked pages. The latent semantic information that causes the improvement in KEA is expected to be transferable to other domains, particularly Web searching. Future work therefore involves (in order of immediacy):

- Further analysis of the distribution of SR in keyphrases and adapting the KEA algorithm to use a more suitable classifier than the naïve Bayes. It is clear that the SR and TFxIDF features are not independent, as the naïve Bayes classifier requires. Furthermore, while phrases with low TFxIDF are generally less likely to be keyphrases, this is not typically the case with SR. Therefore, the independence assumption will, in some cases result in a less accurate classification. Also, the normal density function assumed by KEA may be suboptimal for the continuous features used in this project. Current work therefore involves testing variations on the naïve Bayes classifier (such as the 'flexible' naïve Bayes described in [John *et al.*, 1995]).
- Experimentation with the SR feature is required in order to determine if a more suitable feature or number of features exist. In addition, the SR feature will be more generalised to take into account more distant documents than those directly related to the document in question, perhaps including a link-weighting or spreading-activation mechanism for retrieving the documents.
- The SR distribution will be further analysed and used to adapt a term-frequency-based document retrieval program to the Web and to improve existing web search engines by assigning scores to documents according to the probability that a given search term is a keyphrase of that document. These projects will be applied initially in a domain-specific, supervised learning environment, and ultimately in a more general and universal environment.

The initial results mentioned in this paper lend encouragement to the hope that a system based on hyperlink analysis can make significant improvements to existing Web search technology.

References

[Frank et al., 1999] Frank E., Paynter G.W., Witten I.H., Gutwin C. and Nevill-Manning C.G. (1999) Domain-specific keyphrase extraction *Proc. Sixteenth International Joint Conference on Artificial Intelligence*, Morgan Kaufmann Publishers, San Francisco, CA, pp. 668-673.

[John et al., 1999] John, G.H., Langley, P., (1995) Estimating Continuous Distributions in Bayesian Classifiers. *Proc. Eleventh Conference on Uncertainty in Artificial Intellience,* Morgan Kaufmann Publishers, San Mateo.

[Salton, 1988] Salton, Gerard. Automatic text processing: the transformation, analysis, and retrieval of information by computer, Reading, Mass.Wokingham : Addison-Wesley, 1988.

[Witten et al., 1999] Witten I.H., Paynter G.W., Frank E., Gutwin C. and Nevill-Manning C.G. (1999) KEA: Practical automatic keyphrase extraction. *Proc. DL '99*, pp. 254-256. (Poster presentation)