IRM PRESS
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.irm-press.com
ITB10755

Chapter 9

Word Weighting Based on User's Browsing History

Yutaka Matsuo,
National Institute of Advanced Industrial Science & Technology, Japan

Abstract

This chapter presents discussion of word weighting algorithms in user modelling and adaptive information systems. We specifically address two types of user interest: (1) broad and consistent interest; and (2) narrow, spot interest. A user's consistent interests can be modelled utilising the user's information access history; a user's spot interests can be determined based on that. We developed a word-weighting algorithm to measure the user's spot interest. The information access history of a user is represented as a set of words. It is considered to be a user model. This method weights words in a document according to their relevancy to the user model. The relevancy is measured by the biases of co-occurrence, called the Interest Relevance Measure, between a word in a document and words in the user model. The future methodology of word weighting is described herein while demonstrating our approach.

Introduction

Many information support systems include natural language processing (NLP) techniques, such as document (including Web page) retrieval, summarisation, recommendation, and so on. Those techniques usually use a word-weighting algorithm. For example, to retrieve documents that match a user's query, the system must be able to find documents containing some words (derived from a user's query), and order them based on the weight of the words in each document. A word is important in some documents and not in other documents. If the word is the subject of the document, the weight should be high. Conversely, the weight of the word should be low if the document just mentions the word because of some slight relevance to its subject. Therefore, each document must be indexed by appropriate words using a word-weighting algorithm to retrieve appropriate documents. In summary, a system must extract sentences (or phrases) which have high weight value. That value is typically determined by summing up the weight of words in the sentence and some additional weight (Mani, 2001).

Word weighting research dates back to the 1950s. Since that time, many studies have addressed this topic (Luhn, 1957; Sparck-Jones, 1972; Nagao, Mizutani, & Ikeda, 1976). A famous system was developed by Salton, called the SMART system (Salton & Yang, 1973; Salton, 1989). Currently, the most major algorithm of word weighting is *tfidf;* it is based on statistics of word occurrence in a target document and a corpus. It is demonstrably effective in many practical systems (Pretschner & Gauch, 1999).

Although *tfidf* is a robust and useful approach, it presents limitations from the viewpoint of a user model and adaptive information systems. *Tfidf* depends on a document and a corpus, not on a user. If a document or a corpus changes, the weight is affected, but it is always the same to all users, independent of whether a user has interest or expertise in that area or not. A word that is important to one user is sometimes not important to others. The newspaper article "Suzuki hitting streak ends at 23 games" provides an illustrative example. Ichiro Suzuki is a Japanese-born Major League Baseball player who was recognised as the MVP and Rookie of the Year in 2001. A user who is greatly interested in Major League Baseball would be interested in a phrase such as "hitting streak ends" because the user would know that Suzuki was achieving the longest hitting streak in the majors in that year. On the other hand, a user who has no interest in Major League Baseball at all would note the words "game" and "Seattle Mariners" as informative words, because those words would indicate that the subject of the article was baseball. Such knowledge would be sufficient for that user. In this sense, important words differ for each reader.

Current systems use the weight of words to represent a document profile (usually as a *tfidf* vector) and to compare a document profile with a user profile. Such weighting is a reasonable approach to measure the similarity between a user profile and a document because the system need not re-calculate document profiles for each user; it only has to change the user profile. However, considering future adaptive information systems, this approach is insufficient for two reasons: