Linkage Discovery with Glossaries

L

Richard S. Segall
Arkansas State University, USA

Shen Lu
SoftChallenge LLC, USA

INTRODUCTION

This chapter discusses the topic of linkage discovery with glossaries and reviews the related work of others in this area. Linkage discovery has been proven to be a useful method of relating sections of text, themes, and subtopics and has been enhanced with the use of glossaries for which developments and experimentation are as discussed in this chapter.

BACKGROUND

With the development of computer technology and Internet, the electric publication can do much more than only emulate the printed publication in cheaper and more transportable forms like Web pages and electric files. Electronic publications also have the potential to enhance the reading process itself through the identification of new ways to retrieve, index, and search information throughout the entire proceeding, for knowledge discovery in related text. Latent Semantic Analysis (LSA) can be used to implement those functions and to discover knowledge from text with a general mathematical learning method without knowing prior linguistic or perceptual similarity knowledge.

Latent Semantic Analysis (LSA) is a Natural Language Processing (NLP) technique that is based on similarity of words but not grammatical or syntactical structure and extracts knowledge through the similarity of individual words. The motivation of LSA in terms of psychology is that people learn knowledge only from similarity of individual words taken as units, not with knowledge of their syntactical or grammatical function. LSA assumes that the dimensionality of the context in which all of the local words are represented is of great importance and the reduction of dimensions of the observed data from original text to a much small but still large number can improve human cognition. LSA consists of two steps:

1. Represent the text as a matrix in which each row is a unique word and each column is a text message or other context. Each cell contains the frequency of the word in column of the corresponding passage. The frequency of the cell entry is weighted by a function that expresses both the importance of the word in the particular passage and how much information the word has in general.
2. LSA applies Singular Value Decomposition (SVD) to the matrix.

This chapter pertains to the operation of inserting the definitions of the terms into glossaries of those words of an article. In order to discovery linkage between different sections, one needs to have as much specific knowledge (especially meaningful words) as possible from the text. We then use the domain knowledge to improve the accuracy of the linkage discovery from the context.

However, with many different topics in one book, all of the text in one book is not enough to train models to discover knowledge from different domain areas. Also, domain knowledge is hidden in general terms and is hard to highlight and extract. In this chapter, the research pertains to combined domain knowledge given in the form of domain glossaries for a specified discipline

DOI: 10.4018/978-1-4666-5202-6.ch128

with the text in a book to specify the meanings of different terms and then find similar sections. Experimental result have showed us in Lu et al. (2011) and Lu et al. (2012) and by other investigators that, by combining glossaries with the text, we can extract more meaningful words from the text and then link similar sections together.

Documents contain terms from the corresponding glossaries in the same domain areas. Glossaries include all of the domain knowledge from different areas, which are described by the significant terms and their corresponding definitions. In text mining, one of the issues is how to extract significant terms from the text. All of the terms associated with domain knowledge are distributed everywhere in an article and are mixed with general words which have nothing to do with the domain knowledge.

Latent Semantic Analysis (LSA) can provide the meanings of the terms based on the context. However, one article cannot include all of the domain knowledge and the definition extracted from the context where the term appears in that article is not accurate. But, in glossaries, all of the terms are defined clearly. In Lu et al. (2011) and Lu et al. (2012), we manually put the definitions of the terms in glossaries to those words in an article and use those definitions to improve the accuracy of the background knowledge we can extract from the context. In this way, we can define meaningful words and use them to decide the theme of the corresponding sections.

Ferret (2002) presented a method, called TOPICOLL, for using collocations for topic segmentation and link detection. Figure 1 illustrates the automation of the algorithm of Ferret (2002) for detecting topic shifts.

TOPICOLL Algorithm

Parameters:

1. **State:** {NewTopicDectection, InTopic, EndTopicDetection, OutOfTopic}
2. **Sim:** {True, False} // If the context of the focus window and the context of the current segment is similar, Sim is True; otherwise, Sim is False.
3. **ConfirmNb:** Integer // the number of successive positions.
4. **Tconfirm:** Constant // the threshold of successive positions (Box 1).

Figure 1. Automation for topic shift detection (Ferret, 2002)