Zhang Caigui, German Yuliya Olegovna

METHOD OF SEMANTIC BLOCK DEFINITION IN TEXT

This article introduces the algorithm of semantic block definition in text based on vector space model. Fundamentals of vector space modeling are introduced. Common performance evaluation criteria for text retrieval systems are analyzed.

INTRODUCTION

Natural Language Processing (NLP) is a branch of the field of Computer Science and Artificial Intelligence that aims to allow computers to understand, generate and process human language. In NLP, semantic block definition in text is an important subfield that deals with the ability to find relevant information quickly and accurately in large amounts of textual data.

The main goal of semantic block definition in text in Natural Language Processing is to find the most relevant document or text to a given query. This technique plays an important role in various applications such as search engines, document management systems, knowledge bases, question and answer systems, etc.

In this paper, we will introduce the algorithm based on vector space model for semantic block definition in text.

I. VECTOR SPACE MODEL

In the information retrieval model a document is described by a representative set of words (called index terms). It is usually necessary to extract all the words contained in the set of texts to be processed. The set of all words $T = \{T_1, T_2, ...T_M\}$, where M denotes the number of words contained in the text collection, and M generally increases as the text collection keeps changing. Generally after preprocessing, the words with obvious iconic roles in the document are used as index items. For the initial document $d_j = t_{j1}t_{j2}...t_{jNj}$, where N_j is the document d_j contains the number of words, after preprocessing $d'_j = t_{j_1}t_{j_2}...t_{j_{N_j'}}$, where $N_{j'}\leq\mathrm{N}_j$, preprocessing can be very good to reduce the amount of computation. Representing documents as vectors of indexed item weights is the most common way, which is the vector space model.

II. CORRELATION CALCULATION

The vector space model assigns weights to the index items of a document (or query). The document is represented as a vector of weights $W_j = \langle\ W_{1j}, W_{2j}, ...W_{Mj}\rangle$ where W_{ij} denotes the weight of the index term ti in the document d_j. W_{ij} is computed using the TFIDF weighting strategy, and the specific formula can be expressed as: $W_{ij} = (1 + \log(tf(t_i, d_j))) \cdot \left(\log\left(1 + \frac{N}{df(t_i)}\right)\right)$

Where $tf(t_i, d_j)$ is the number of occurrences of the word t_i in the document d_j; N is the number of texts to be clustered; $df(t_i)$ is the number of documents containing the word t_i. At the same time, the query Q needs to be represented as a vector of weights to calculate the similarity between the query and the documents. The query is represented as $Q = \langle\ W_{1q}, W_{2q}, ...W_{Mq}\rangle$. The size of W_{ij} is proportional to the number of occurrences of t_i in document d_j and inversely proportional to the number of occurrences of t_i in the entire collection of text. The formula for similarity is expressed as: $Sim(Q, d_j) = \frac{\sum_{k=1}^{M} W_{ki} * W_{kj}}{\sqrt{\left(\sum_{k=1}^{M} W_{ki}^2\right)\left(\sum_{k=1}^{M} W_{kj}^2\right)}}$

SUMMARY

Text retrieval in natural language processing has made great progress, but still faces some challenges like current text retrieval and text search algorithms cannot fully understand the semantics of natural language, and therefore cannot handle complex queries.

In the future, text retrieval and text search in natural language processing will face the following development trends:

1 Deep learning-based models: with the development of deep learning technology, text retrieval and text search models based on deep learning will be more widely used.

2 Semantic search: future text search will pay more attention to user needs and provide more accurate search results through semantic understanding.

References

1. Arvind Arasu, Junghoo Cho, Hector Garcia M. Searching the Web. ACM Transactions on Internet Technology. Auguest 2001, 1(1): 43
2. Sui Zhifang, Chen Yirong, Hu Junfeng. The research on the automatic term extraction in the domain of information science and technology. Institute of Computing Linguistics, Peking University.
3. Cay S Horstmann, Gary Cornell. Java 2 core technology. Beijing: Machinery Industry Press, 2003.

Zhang Caigui, master's student of the Faculty of Information Technology and Control of BSUIR,zhangcaigui309@gmail.com

German Yuliya Olegovna, PhD, Associate Professor of Information Technologies in Automated Systems Department, Faculty of Information Technology and Control of BSUIR, jgerman@bsuir.by