一种基于软构件描述文本信息抽取的检索方法

韩忠愿，谢 丹
（南京财经大学 信息工程学院，江苏 南京 210046）

摘 要：通过对目前应用广泛的软构件检索技术的研究，提出了一种基于软构件描述文本信息抽取的检索方法。该方法利用中文分词技术和向量空间模型中“词频与倒文档频度”算法抽取关键词，通过《知网》语义相似度，计算用户需求与可重用软构件的匹配度，实现了对软构件的语义检索，能实现模糊查询，具有一定的张弛能力。

关键词：构件检索；向量空间模型；知网；语义相似度；信息抽取
中图分类号：TP311 **文献标识码**：A **文章编号**：1674-7720(2013)02-0001-03

A retrieving method for software components based on text information extraction

Han Zhongyuan, Xie Dan
(Institute of Information Engineering, Nanjing University of Finance & Economics, Nanjing 210046, China)

Abstract: Based on the study of present widely-used software component query technology, a retrieving method for the software components based on text information extraction is proposed in this paper. This method makes full use of the Chinese word segmentation and the keyword extraction by the "term-frequency inverse-document-frequency" algorithm in vector space model, approaching the user query with software components through the words semantic similarity based on《HowNet》, so as to realize the purpose of software component semantic retrieving and support fuzzy query with the property of flexibility.

Key words: component retrival; VSM; HowNet; semantic similarity; information extraction

随着软件开发规模的增大，软件构件技术被认为是解决软件危机的有效途径，基于构件的软件开发 CBSD (Component Based Software Development)[1] 成为有效提高软件生产率、缩短软件产品交付时间和提高软件质量的新方法。

传统的软构件的检索方法[2]主要有三种：基于外部索引的检索、基于内部静态索引的检索和基于内部动态索引的检索。其中以构件的刻面表示以及在此基础上的构件检索技术已得到软件复用界的重视和应用[3]。著名的 REBOOT 构件库[4]提出了可重用软件构件基于刻面的分类检索方案。国内的青鸟构件库[5]采用以刻面分类为主、多种分类模式相结合的方法对构件进行分类描述。

传统的基于关键字或刻面描述的软件构件的检索由于缺少特定领域语义信息，使得用户在查询所需要的构件时，有时很难对构件的各个刻面作出准确的描述，因此在查准率和查全率上存在不足。准确地理解用户的查询请求是构件检索的一个重要问题，本文针对与软构件如影随形的自然语言描述，提出一种基于软构件描述文本信息抽取的检索方法。该方法采用自然语言描述软构件的实现，并由系统利用自然语言处理技术抽取软构件特征信息和需求的特征信息，然后利用特征匹配和《知网》词汇语义相似度计算获得候选的结果。

1 软构件检索系统体系结构

有效的构件检索机制能够降低构件查找和理解的成本，检索方式对构件描述和用户查询的依赖是本文研究的主体部分。本文设计了基于文本描述的软构件检索系统体系结构，如图 1 所示。其各部分功能如下：

(1) 软构件文本描述主要是将系统数据库中有关软构件的文本描述信息提取出来进行自然语言处理，并将处理返回的结果存储起来；主要负责与用户交互，为用户提供查询接口，通过用户输入生成查询条件，并将满足条件的软构件信息返回给用户。

(2) 自然语言处理模块主要是将数据库的软构件文本描述信息集合在一起，通过 ICTCLAS 分词技术获得带标注的分词结果，并根据 VSM 中 TFIDF 的计算方法为每个软构件描述文本提取特征项并存储；

(3)检索模块分为两种方式：一种是将用户查询的特征与抽取出的软构件特征项通过《知网》词汇语义相似

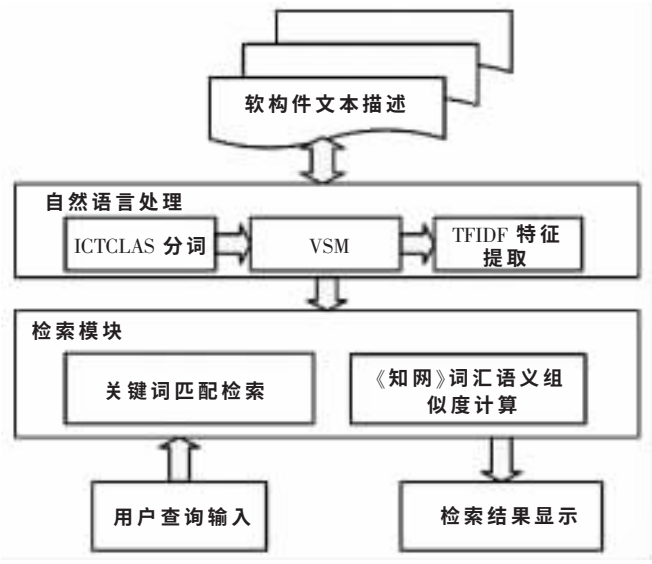

图 1 软构件检索系统的体系结构

度计算来获取查询结果，此种方法主要实现了软构件的语义检索，是本文研究的重点；另一种是用户查询的特征与软构件特征项之间的匹配检索。

这种层次结构的体系模式将各模块的功能相互独立，有利于系统的维护与扩展，确保了系统的稳定性和可维护性。

2 软构件检索实现分析

检索实现是本文研究的重点，尤其是实现软构件的语义检索。通过上面软构件检索系统的体系结构图可以看出，自然语言处理部分是实现语义检索的基础，自然语言处理的准确度直接影响到检索结果的查全率和查准率。

下面简单介绍 ICTCLAS 汉语分词系统和 VSM 的研究现状，并详细介绍语义检索的实现过程。

2.1 ICTCLAS 汉语分词简介

分词系统[6]ICTCLAS(Institute of Computing Technology,Chinese Lexical Analysis System)是由中科院计算所的张华平、刘群所开发的一套获得广泛好评的分词系统。它先通过层叠形马尔可夫模型 CHMM(Hierarchical Hidden Markov Model)进行分词，通过分层，既增加了分词的准确性，又保证了分词的效率。ICTCLAS 分词速度单机 500 KB/s，分词精度 98.45%，是世界上最好的汉语词法分析器，并且在国内 973 专家组组织的评测中获得了第一名。

2.2 向量空间模型

向量空间模型 VSM(Vector Space Model)由 Salton 等人于上世纪 60 年代末提出，并成功应用于著名的 SMART 系统，是目前最为成熟且应用最为广泛的文本表示模型之一[7]。它把对文本内容的处理简化为向量空间中的向量，用“词频与倒文档频度”TFIDF(Term-Frequency Inverse-Document-Frequency)[8] 进行特征项赋权值，来表征某个特征项对该文本内容的重要程度。其中 TFIDF 将一个特征项在某个文档中的重要性和在整个文档数据全集中的重要性结合起来，成为一个统一的度量值。它说明一个在单个文档中频度很高，而在整个数据全集中频度很低的词是更加重要的词。

本文在自然语言处理过程中对文本关键词的抽取正是提取 VSM 中 TFIDF 值较高的特征项，将通过此方法获得的所有特征项按权值大小排序，提取满足阈值或一定数目的最优特征作为最终表达该文本特征的特征项集。

2.3 《知网》词汇语义相似度计算

《知网》(HowNet)[9] 是一部比较详尽的语义知识词典，是一个以汉语和英语词义所代表的概念为描述对象，以揭示概念间及概念所具有的属性间关系为基本内容的常识知识库。概念与义原是《知网》中的两个主要概念。每一个词可以表达为几个概念，每个概念又可由若干个义原来描述。

对于两个汉语词语 W_1 和 W_2，如果 W_1 有 n 个概念，S_{11}，S_{12}，$\cdots$，S_{1n}；W_2 有 m 个概念，S_{21}，S_{22}，$\cdots$，S_{2m}，则《知网》规定，W_1 和 W_2 的相似度是各个概念的相似度之最大值，即：

$$\mathrm{Sim}(W_1,W_2)=\max_{i=1,2,\cdots,n;j=1,2,\cdots,m}\mathrm{Sim}(S_{1i},S_{2j}) \tag{1}$$

如此，就将两个词语之间的相似度问题归结到了两个概念之间的相似度问题。由于义原是描述一个概念的最小意义单位，所以义原的相似度计算是概念相似度计算的基础，概念相似度是由提取到的义原的相似度加权平均得到的。

假设两个义原在同一个层次体系中的路径距离为 d，可以得到这两个义原之间的语义距离：

$$\mathrm{Sim}(p_1,p_2)=\frac{\alpha}{d+\alpha} \tag{2}$$

其中，p_1 和 p_2 表示两个义原；d 是 p_1 和 p_2 在义原层次体系中的路径长度，是一个正整数；α 是一个可调节参数，一般取值 1.6。

以上是《知网》词汇语义相似度的计算方法，是本文的一个重要部分，精确的词汇匹配度是下一步检索的基础工作。

2.4 检索模块

通过抽取软构件文本描述特征项来实现基于语义的检索是本文研究的重点。通过对相似度计算模块得到的数据进行处理分析，是实现检索的关键步骤，其主要处理流程如图 2 所示。

假设用户查询关键词集合为 $Q\{K_1,K_2,\cdots,K_m\}$，某一软构件的文本描述向量空间模型的特征项表示为集合 $D_i\{T_1,T_2,\cdots,T_n\}$，其中 $T_j\{j=1,2,\cdots,n\}$ 为经过自然语言处理的描述该构件的特征项。

一般将两个集合中的特征项两两比较得到的相似度的平均值作为它们的相似度，如此一个集合任意两个特征项之间的相似度都为 1，集合才能与它本身 100% 相似。本文采用以下算法为这两个集合进行相似度计算：

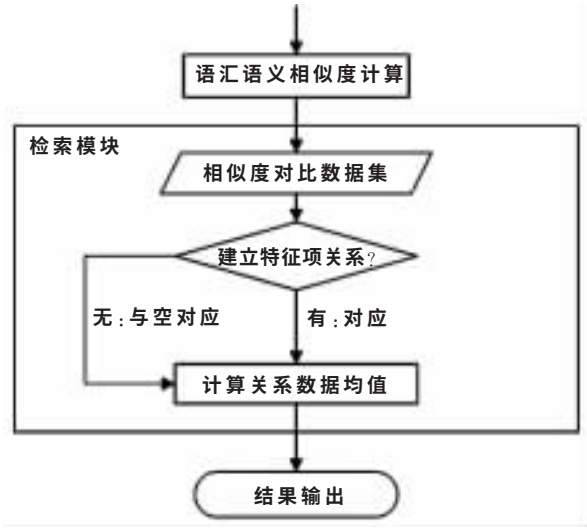

图 2 检索模块数据处理流程

(1)利用《知网》词汇语义相似度,将 Q 中每个关键字与 D_i 中每个特征项进行相似度计算,如图 3 所示。得到 Term_Sim {Sim (K_1,T_1),Sim (K_1,T_2), …,Sim (K_i,T_j), …,Sim (K_m,T_n)}为相似度值集合,共 $m\times n$ 个数据。

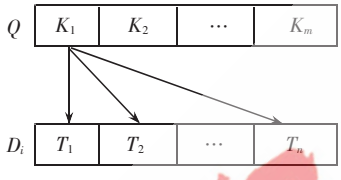

图 3 词汇相似度对比

(2)将相似度值中最大的值所对应的 K_i 和 T_j 建立对应关系。

(3)将包含 K_i 和 T_j 的相似度值从 Term_Sim 中删除。

(4)重复(2)和(3),直到所有的相似度值都被删除。

(5)没有建立起对应关系的关键字或特征项与空对应。

(6)将包含 K_i 的相似度值取算术平均值。

把上面得到的平均值作为用户查询与软构件之间的相似度度量值,将满足阈值的软构件信息按照相似度值的递减顺序输出。

3 实验结果

根据以上描述,实现了在 ERP 领域软构件的检索,检索结果如图 4 所示。

实验从 ERP 软构件描述数据库中抽取出相似度较高的软构件作为候选结果输出。其中,“成本管理”经过 ICTCLAS 分词、VSM 处理得到的关键词是:“成本”、“产品”、“计算”等,与用户检索关键词“成本”、“分析”比较,相似度值是 72.22%。在查询结果中点击相应的项目,会详细显示对构件的描述,可以帮助用户更清晰地了解该构件的信息,从而从候选结果中选择符合要求的软构件。

本文提出了一种基于文本信息抽取的软构件检索方法,并对软构件检索系统的体系结构、功能模块进行了详细介绍,优化了关键字集合相似度计算;并且针对传统软构件检索中语义缺失的缺点,实现了对软构件的语义检索的目的,有利于进行基于软构件的软件开发。另外,本系统还有尚待改进的地方,例如:扩充分词词典,保证领域术语的完整性;增加软构件的图形描述,实现多功能检索等,这些问题也是下一步研究工作的重点。

图 4 检索结果

参考文献

[1] BROWN A W, WALLNAU K C. The current state of CBSE[J]. IEEE Software, 1998, 15(5):37-46.

[2] 刘韬,范菁,熊丽荣.构件的检索技术研究及其在信用领域构件库中的应用[D].杭州:浙江工业大学,2008.

[3] 舒远仲,陈志勇,彭晓红,等.基于刻面分类描述的构件检索方法研究[J].计算机工程与科学, 2010,32(11):156-160.

[4] MOREL J M, FAGET J. The REBOOT environment[C].In: Prieto-Diaz R,Frakes WB eds.Processdings of the 2nd International Workshop on Software Reusability Advances in Software, Lucca:IEEE Computer Society Press, 1993:80-88.

[5] CHANG J C, LI K Q,GUO L F,et al. Representing and retrieving reusable software components in JB(Jadebird) System[J]. Electronica Journal,2000,28(8):20-24.

[6] ICTCLAS 分词系统研究[EB/OL].(2010-08-24).http://wenku.baidu.com/view/2eeb4afff705cc175527093f.html.

[7] 杨小平,丁浩,黄都培.基于向量空间模型的中文信息检索技术研究[J].计算机工程与应用,2003(15):109-111.

[8] 王晓龙,关毅.计算机自然语言处理[M].北京:清华大学出版社,2005.

[9] 刘群,李素建.基于《知网》的词汇语义相似度计算[C].台北:第三届汉语词汇语义学研讨会论文集,2002:59-76.

(收稿日期:2012-09-16)

作者简介:

韩忠愿,男,1963 年生,博士,硕士生导师,主要研究方向:软件工程,软构件技术,自然语言处理。

谢丹,女,1989 年生,硕士研究生,主要研究方向:软件工程,软构件技术,自然语言处理。