面向视频场景内容检索的文本解析工具设计与实现*

吴洁明,周正喜,史建宜
(北方工业大学 信息工程学院,北京 100144)

摘 要:在足球赛事视频的应用背景下,分析了面向视频场景内容检索的文本解析关键技术,设计并实现了面向视频场景内容检索的文本解析工具。该工具利用中文分词技术分割自然语言文本,通过汉语语法规则提取关键词,采用加权算法对关键词排序,并将关键词映射到知识表达集,从而获得关键词的语义信息,完成文本解析。实验结果表明,该工具能够满足自然语言文本的视频检索需求。

关键词:视频场景内容检索;文本解析;关键词提取;知识表达;关键词映射
中图分类号:TP391 **文献标识码**:A **文章编号**:1674-7720(2012)14-0070-05

Design and implementation of a text retrieval tool based on video content searching

Wu Jieming, Zhou Zhengxi, Shi Jianyi
(Department of Information Engineering, North China University of Technology, Beijing 100144, China)

Abstract: In the background of application for soccer video, the paper analyzes key technologies on the text retrieval tool for video content, and designs and implements a text retrieval tool for video content. Firstly, the tool segment natural language text by using Chinese word segmentation, extract key words by Chinese grammar rules from the segmentation of natural language text, sort keywords by using the weighted algorithm, map the keywords to knowledge expression set to gaining semantic information of keywords and completing text parsing, and finally application of text retrieval tool for video search is showed by using the web page. The experiment results show that the text analysis tool for video content designed and implemented in the paper meets the demand for searching video by natural language text.

Key words: video content retrieval; text retrieval; keyword extraction; knowledge representation; keyword mapping

本文依托“基于视频素材的虚拟场景生成系统及其关键技术研究”课题,图1是该课题的部分框架。

图1 依托课题的部分框架图

通过视频场景内容标注工具对视频帧进行标注,根据知识表达集生成视频的描述标注信息,并且通过视频数据库保存起来。输入的检索文本通过视频场景内容检索工具,检索返回用户所需要的视频。

文本解析模块需面向视频场景内容,基于语义对文本进行解析,以匹配视频语义标注集,从而获得更好的检索效果。本文主要描述了面向视频场景内容检索的文本解析工具的设计思想和实现原理。

1 相关工作

1.1 中文分词

中文分词是将中文字序列按照一定的规则重新组

*基金项目:“十二五”国家科技支撑计划项目(2012BAH04F03)

合成词序列的过程[1]。在对文本进行解析时，词是最小的能够独立活动的有意义的语言成分。没有中文分词，其他一切深入的中文信息处理都无从谈起[2]。

在中文分词方面已经有精确度很高的分词算法和工具，特别是中国科学院计算所推出的ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System)，是由中科院计算所的张华平、刘群所开发的一套分词系统，其汉语分词、未定义词识别、词性标注和人名识别的效果广受好评[3-4]。

1.2 关键词提取

关键词能够以最简洁的形式概括表达文章主体大意，可用来检索海量信息。

目前大部分的关键词提取算法都是基于机器学习的方法。在这些算法中，同一篇文档中的同一个词在不同的地方或许有不同的意思，例如“苹果”能够表示水果的一种或者苹果产品的意思。同样地，不同的词能够表示相同的意思。这些现象产生的原因在于词汇层面(代表意思的词)和概念层面(意思本身)的差别，这样将会导致关键词提取的不准确[5]。

国内外对于文本关键词的提取研究主要是针对文档、Web页面等大文本，而本系统针对的是小文本。本文采取基于汉语语法规则和中文分词系统分词的词性，进行关键词的提取。如对于主体而言，在经过分词后，词性被确定为“nr”或“r”(如“贝克汉姆”词性为“nr”，“他”词性为“r”)，则该词语为主体；如对于行为而言，词性被确定为“v”(如“射”词性为“v”)，则该词语为行为的一部分。

1.3 加权算法

一般的文本具有有限的结构或根本没有结构，文本解析的目的是抽取出该文本的结构特征并用结构化的形式保存。本文选用经典的TF-IDF算法表示特征的权重。

在一份给定的文件里，词频(TF)指某一个给定的词语在该文件中出现的次数，对于在某一特定文件里的词语 t_i 来说，它的重要性为：

$$tf_{i,j}=\frac{n_{i,j}}{\Sigma knkj}$$

其中，$n_{i,j}$ 是该词在文件 d_j 中的出现次数，而分母是在文件 d_j 中所有字词的出现次数之和。逆向文件频率(IDF)是一个词语普遍重要性的度量。某一特定词语的IDF为：

$$idf_i=\log\frac{|D|}{|d:t_i\in d|}$$

其中，$|D|$ 为语料库中的文件总数，$|d:t_i\in d|$ 表示包含词语 t_i 的文件数目(即 $n_i\neq 0$ 的文件数目)。再由

$$tfidf_{i,j}=tf_{i,j}\times idf_i$$

计算出某一特定文件内的高频率词语。过滤常见的词语，保留重要的词语[6]。

1.4 关键词映射技术

视频检索文本经过中文分词后，提取了关键词及其权重，并且生成了视频检索结构，但是，其与知识表达集中的信息不一定完全符合，需要将关键词映射到知识表达集，然后进行检索。例如输入的视频检索文本“范尼斯特鲁伊小禁区内右脚射门”经过关键词提取与权重排序模块之后，生成了关键词“范尼斯特鲁伊”，但是知识表达集里没有定义“范尼斯特鲁伊”，而只是定义了“人”或者是“运动员”，所以，要依据一定的技术(如数据字典、人名识别等)将“范尼斯特鲁伊”映射为知识表达集中的“人”或者“运动员”，再生成到知识表达集的检索结构，供检索算法调用。在这部分工作中，涉及到三方面的问题：

(1)实例到概念的映射。如“罗纳尔多”是“运动员”这个概念的一个实例，则在检索过程中，应该将“罗纳尔多”映射成为“运动员”。

(2)同义词匹配。如果检索语句中有“门将”关键词，而知识表达集中只有相对应的“守门员”这个本体，则应该将二者匹配起来。

(3)未登录词映射。当检索语句中有知识表达集中没有涉及到的概念或本体时，则需要做关键词映射，例如将“球员”映射为“运动员”。

2 面向视频场景内容检索的文本解析工具

2.1 总体设计

面向视频场景内容检索的文本解析工具主要包括4个模块：中文分词模块、关键词提取模块、关键词权重排序模块和关键词映射模块，如图2所示。输入的视频检索文本经中文分词模块生成词语及词性序列，关键词提取模块生成主体、行为、场景、身体部位等关键词，经过关键词权重排序模块生成带有权值的关键词结构，最后做关键词到知识表达集的映射，优化检索。

图2 文本解析工具结构图

2.2 中文分词模块

对于输入的检索文本，汉语分词系统ICTCLAS能分解出各个成分，并且确定各个成分的词性。如检索文本“范尼斯特鲁伊小禁区内右脚射门”，分词的最终结果为“范尼斯特鲁伊/nr 小/h 禁区/n 内/f 右/f 脚/n 射门/v”。其中，“nr”表示“人名”，“h”代表“前缀”，“n”表示“名词”，“f”表示“方位词”，“v”表示“动词”。

2.3 关键词提取模块

本文应用背景主要是对足球运动视频的检索。关键词

提取模块基于中文分词模块的切分句子和分出各个成分词性的分词结果，通过汉语语法规则，提取能表现足球运动视频帧的主体、行为、场景和身体部位等关键词。

设检索文本中第 i 个词称为词 i，其词性用 partOfSpeech[i]表示。

2.3.1 主体关键词提取

(1)当 partOfSpeech[i]为“nr”(人名，如“贝克汉姆”)或“r”(代词，如“我”)时，词 i 是主体；

(2)当 partOfSpeech[i]为“pbei”(被动词)，且 partOfSpeech[$i+1$]为“n”时，词 $i+1$ 是主体；

(3)当 partOfSpeech[i]为“pbei”，且 partOfSpeech[$i+1$]为“v”时，被动词后应有一个主体，主体类型(运动员、守门员，裁判)由“v”的类型发出者决定，此时的主体为补充主体。

2.3.2 行为关键词提取

行为关键词主要涉及到动词及其宾语，本文中，行为主要是动词+宾语。

当 partOfSpeech[i]为“v”或“vg”或“vn”或“vd”或“vJudge_word”(自定义词性)，且词 $i+1$ 词性为“v”或“n”或“ng”，且词 $i+2$ 词性为“v”或“n”时，行为=词 i+词($i+1$)+词($i+2$)。如果词 $i+2$ 词性不为“v”且不为“n”时，行为=词 i+词($i+1$)。如果词 $i+2$ 词性不为“v”且不为“n”，且词 $i+1$ 词性不为“v”且不为“n”且不为“ng”时，行为=词 i。

2.3.3 场景关键词提取

(1)当 partOfSpeech[i]为“p”，且 partOfSpeech[$i+1$]为“n”或“s”时，词 $i+1$ 是场景；

(2)当 partOfSpeech[i]为“n”，且 partOfSpeech[$i+1$]为“f”，且词 $i-1$ 的词性为“a”时，场景=词($i-1$)+词 i+词($i+1$)。如果 partOfSpeech[$i-1$]不为“a”，则场景=词 i+词($i+1$)；

(3)当 partOfSpeech[i]为“f”，且 partOfSpeech[$i+1$]为“q”，则场景=词 i+词($i+1$)；

(4)当 partOfSpeech[i]为“scene_word”(自定义词性)，则词 i 是场景。

2.3.4 身体部位关键词提取

(1)当 partOfSpeech[i]为“pyong”，则词 i 是身体部位；

(2)当 partOfSpeech[i]为“f”，且 partOfSpeech[$i+1$]为“n”，则身体部位=词 i+词($i+1$)；

(3)当 partOfSpeech[i]为“f”，且 partOfSpeech[$i+1$]为“q”，则身体部位=词 i+词($i+1$)；

(4)当 partOfSpeech[i]为“bodyPart_word”(自定义词性)，则词 i 是身体部位。

2.3.5 关键词提取模块的实现

关键词提取模块为了存取 4 类关键词，涉及了两种数据结构。第一种是存储分词后各词的名称及词性的一个二维数组，具体为 String deletedSymbolResult[][]=new String[TEXTNO][2]，其中，TEXTNO 表示一次分词过程中可能涉及到的最多的词语数目，deletedSymbolResult[i][0]存取词语的名称，deletedSymbolResult[i][1]存取词语的词性。第二种数据结构是类 KeywordStruct、subject、action、scene 和 bodyPart，分别存取主体、行为、场景和身体部位关键词。由于每个主体的行为可能有多个，因此，行为 action 以数组形式存取：public String action []=new String [127]。因为关键词提取模块是在中文分词基础上进行，所以，要对中文分词进行优化以将分词结果存于合适的数据结构中，加载自定义词典以弥补原有词典的不足，然后进行 4 类关键词的提取。关键词提取模块对关键词的提取分为以下几个部分：

(1)中文分词处理结果的优化，包括删除分词结果中标点符号和以空格和“/”分隔符分割的分词结果，抽取详细分词信息这两部分。中文分词的结果以字符串 wordSegResult(如“范尼斯特鲁伊/nr 小/h 禁区/n 内/f 右/f 脚/n 射门/v”)表示，调用 split("\\s")和 split("/")方法，可以以空格和“/”分割分词结果，并且将分割的结果存入 deletedSymbolResult 中，分词结果中标点符号的删除以循环 deletedSymbolResult 数组、剔除标点的方法实现。

(2)加载自定义字典，重新定义部分词语的词性。在中文分词处理结果优化的基础上，调用类 KeywordStruct 中 plusDictionary()方法，包括将“守门员”这类普通名词“n”重新定义为能识别出其主体特征的“nr”词性，将“黄牌警告”重新定义为“vJudge_word”这类能识别出主体发出者为“裁判”的词性等。

(3)进行主体、行为、场景和身体部位关键词的识别和提取，包括存储关键词信息的类 KeywordStruct 实例化和各个关键词的存取。本文考虑到输入的视频检索文本中可能对应多个主体，将类 KeywordStruct 实例化为一个字符串数组变量。数组的每个元素对应一条信息(主体、行为、场景和身体部位等信息)。KeywordStruct keyword[]=new KeywordStruct[3]，下面以代码结合实例“何塞·保罗·格雷罗经过连续传球精密配合丹尼·墨菲助攻西蒙·戴维斯破门得分”说明。调用 GetKeywords 类的 insertSubject()方法，调用 Getter 和 Setter 方法存取主体，Keyword[0].setSubject(“何塞·保罗·格雷罗”);Keyword[1].setSubject(“丹尼·墨菲”); Keyword[2].setSubject(“西蒙·戴维斯”)。调用 key.insertAction()方法，调用 Getter 和 Setter 方法提取行为。如对于主体“何塞·保罗·格雷罗”，存入行为 Keyword[0].setAction(0,“传球”); Keyword[0].setAction(1,“配合”); 对于主体“丹尼·墨菲”，存入行为 Keyword[1].setAction(0,“助攻”)。调用 key.insertScene()方法和 key.insertBodyPart()方法相应地获得对应主体行为发生的场景和该行为发生时所用的身体部位。

2.4 关键词权重排序模块

关键词权重排序主要是为了确定 4 类关键词对于输入的视频检索语句的重要程度，即检索时与视频的相关度。

设计存取主体、行为、场景和身体部位这4类关键词个数的类Item，其中subjectNo、actionNo、sceneNo和bodyPartNo分别表示4类关键词的个数，sub_act_sce_bod_No表示一个句子中主体、行为、场景和身体部位关键词总个数，allSentencesNo表示总的句子数。

为了计算逆向文件频率IDF，首先调用countWordFrequency()读取Text4TF-IDF.txt中的测试语句（文本中每一行为一个测试语句），统计allSentencesNo，同时调用tfidf.Item类中的Getter和Setter方法，存取每个句子中subjectNo、actionNo、sceneNo和bodyPartNo。这样，可以得出allSentencesNo和包含某个关键词的句子个数sentenceWithWord [i](i=0表示包含主体的句子总数；i=1表示包含行为的句子总数；i=2表示包含场景的句子总数；i=3表示包含身体部位的句子总数)。由idf [i]=log(allSentencesNo/sentenceWithWord[i])计算出某个关键词的逆向文件频率IDF。词频TF针对某个特定语句。对于输入的测试文本，获取该语句中subjectNo、actionNo、sceneNo和bodyPartNo，通过计算获得各关键词的词频，结合IDF，计算各个关键词的重要性排序。

2.5 关键词映射模块

2.5.1 同义词匹配映射

关键词检索时，若关键词与视频库中标注的视频帧信息不同，则首先与视频库中的知识表达集进行同义词(本体注释)的匹配。在知识表达集中，对于每一个本体，其对应着本体到本体注释(同义词)的一个结构，当关键词与本体名称不匹配时，就利用这个结构去匹配本体注释。

2.5.2 实例到概念的映射

关键词检索时，若关键词与视频库中标注的视频帧信息不同，并且关键词与匹配本体注释也不匹配，则需要查询本体树结构(本体树定义在一个XML文档中)。如图3所示。

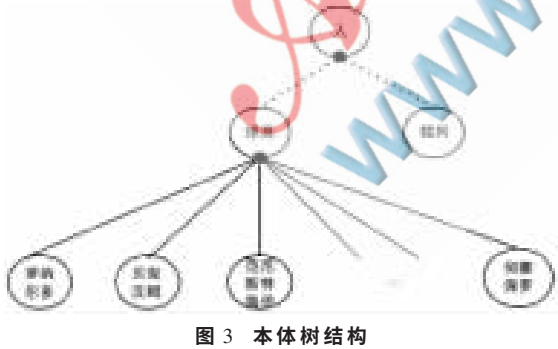

图3 本体树结构

以上的关键词到知识表达集的映射过程可以概括为如图4所示的活动图。首先进行同义词匹配，若匹配成功，则从视频库获取视频的名称、URL等信息，输出视频信息；若匹配不成功，则进行实例到概念的匹配，如果匹配成功，则从视频库里获取并输出视频信息；如果匹配不成功，输出“无对应视频信息”的提示。

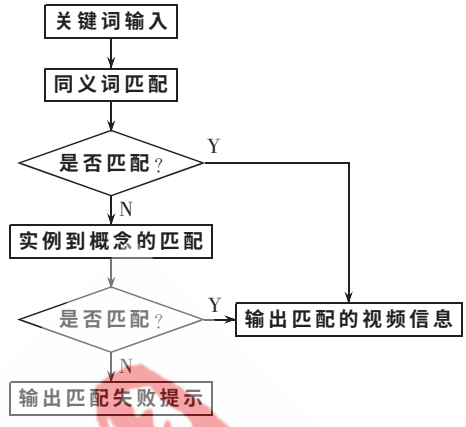

图4 关键词到知识表达集的映射流程图

2.6 文本解析工具

总体上，面向视频场景内容检索的文本解析工具采用B/S架构[7-8]，后台是中文分词模块、关键词提取模块、关键词权重排序模块和关键词到知识表达集的映射模块等组成的Web服务器。在Web浏览器上提交对视频检索文本的中文分词、关键词提取、关键词权重排序、关键词到知识表达集的映射请求，服务器接收到Servlet传来的请求后依次通过中文分词、关键词提取、关键词权重排序、关键词到知识表达集的映射处理，对于中文分词结果、关键词提取结果和关键词权重排序结果，直接以文本形式显示在网页上，对于关键词到知识表达集的映射结果，以检索到的视频输出到网页上的形式间接反映。

当输入的检索文本请求检索页面执行时，如果检索文本为空，则跳转到错误处理页面，返回检索页面继续请求；当检索文本不为空时，跳转到分词结果、关键词提取结果、关键词权重结果显示页面，然后延迟一段时间（如3 s），跳转到视频输出页面，点击视频输出页面上的某个视频（图片或文字链接），则跳转到视频播放页面进行视频播放。

3 实验与分析

下面的实验验证面向视频场景内容检索的文本解析工具的功能，分别包括关键词提取模块、关键词权重排序模块和整体工具的正确性和鲁棒性的检测。

3.1 实验1

输入检索文本经中文分词后，提取面向足球视频背景的主体、行为、场景和身体部位4类关键词，见测试表1。

经过统计可知关键词提取模块在提取关键词时的正确率为93.5%，也可以看到，由于分词模块对某些人名（如“何塞·保罗”）的识别错误，导致提取关键词时，将主体分为两部分，而不是一个主体。总体上，关键词提取模块的关键词提取功能具有可行性。

3.2 实验2

输入检索文本经关键词提取后，通过关键词权重排序模块确定各个关键词在不同大小的测试集中的权重。

在文件test4TFIDF.txt中放入不同数量的测试语句，

表 1 关键词提取模块功能测试表

序号	测试文本	关键词提取结果			
		主体	行为	场景	身体部位
1	范尼斯特鲁伊禁区内射门	范尼斯特鲁伊	射门	禁区内	
2	何塞·保罗禁区外右脚射门	何塞 保罗	射门	禁区外	右脚
3	守门员将球死死抱住	守门员	抱住	禁区外	
4	佐尔坦·格拉射门	佐尔坦·格拉	射门		
5	他在中场传球射门	他	传球射门	中场	
6	埃里克·内维兰德替换下潘特西尔	埃里克·内维兰德	替换		
7	布里德·汉格兰德禁区内头球攻门	布里德·汉格兰德	攻门	禁区内	头
8	佐尔坦·格拉利用对手防守失误，左脚射门得分	佐尔坦·格拉 对手	利用射门得分 防守失误		左脚
9	弗兰克·罗斯特被黄牌警告	弗兰克·罗斯特 裁判	黄牌警告		
10	范尼斯特鲁伊接到他长传球，右脚 射门得分	范尼斯特鲁伊 他	接到射门得分 传球	右脚	

调用 countWordFrequency()方法，统计主体、行为、场景和身体部位 4 类关键词的 IDF 值。为了公平起见，计算 TF 时，设一个句子中 4 类关键词都有且仅有一个，因此，可以用 IDF 代替 TF-IDF。图 5 中横坐标是测试文件中包含的测试语句的个数，纵坐标是 TF-IDF 值。

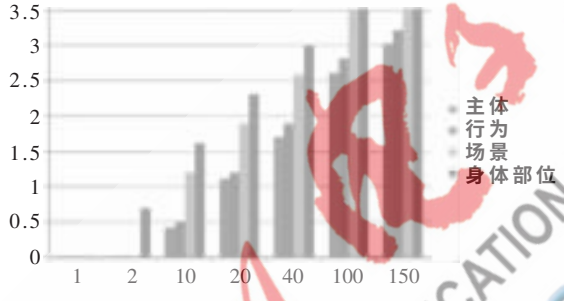

图 5 关键词权重实验结果表

由图 5 可知，每个关键词的 TF-IDF 值都随着文本测试数据数量的增加而增加，对于不同数量的测试数据，涉及身体部位视频的 TF-IDF 最高，其次是场景，主体和行为的 TF-IDF 值最低。说明对于一条视频检索语句，关键词权重由大到小依次是身体部位、场景、行为和主体。

3.3 实验 3

对面向视频场景内容文本解析工具的整体功能进行实验。在主页上输入视频检索文本“范尼斯特鲁伊小禁区内右脚射门”并点击“视频搜索”按钮时，显示了输入的视频检索文本、经过中文分词的结果、关键词提取结果和关键词权重排序的结果。3 s 后，页面自动跳转到视频检索结果页面，显示检索出的视频。

通过实验验证了面向视频场景内容检索的文本解析工具对视频检索文本的验证性处理及错误处理功能，对中文分词处理、关键词提取处理、关键词重要性排序处理的功能，对视频显示和播放的支持。因此，验证了面向视频场景内容检索的文本解析工具的整体功能。

本文设计并实现了面向视频场景内容检索的文本解析工具，该工具包括中文分词、关键词提取、关键词权值排序和关键词到知识表达集的映射模块。特别是关键词提取部分创新性地利用了分词词性与汉语语法规则相结合的处理方式，效果显著。

在中文分词模块，对于非著名球员的名字的识别率存在一定的问题，主要是人名库没有收录；在关键词到知识表达集的映射中，没有考虑第三方面的问题——未登录词映射。下一步的工作将从这两个方面对算法进行改进，以提高分词和检索的准确度。

参考文献

[1] 张素智，刘放美.基于矩阵约束法的中文分词研究[J].计算机工程，2007，33(15)：98-100.

[2] 曹卫峰.中文分词关键技术研究[D].南京：南京理工大学，2009.

[3] 刘群，张华平.ICTCLAS_ 简介[EB/OL].(2008-12-20).[2012-03-10].http://ictclas.org/ictclas_introduction.html.

[4] Zhang Huaping，Liu Qun.Model of Chinese words rough segmentation based on N-shortest-paths method[J].Journal of Chinese Information Processing(in Chinese)，2002，16(5)：3-9.

[5] 方俊，郭雷，王晓东.基于语义的关键词提取算法[J].计算机科学，2008，35(6)：148-151.

[6] 维基百科. TF-IDF[EB/OL].(2010-03-20).[2012-03-10].http://zh.wikipedia.org/wiki/TF-IDF.

[7] ECKEL B.Thinking in Java[M].北京：机械工业出版社，2007：1-600.

[8] HUNT A，THOMAS D.The pragmatic programmer[M].Boston，Massachusetts：Addison-Wesley Professional，2004：3-250.

(收稿日期：2011-12-20)

作者简介：

吴洁明，女，1958年生，教授，主要研究方向：软件工程。

周正喜，男，1986年生，硕士研究生，主要研究方向：软件组件，自然语言处理。

史建宜，女，1988年生，硕士研究生，主要研究方向：自然语言处理，软件工程。