一种摘要中隐含的知识片段的挖掘方案

戴 璐 丁立新 薛 兵
(武汉大学软件工程国家重点实验室 武汉 430072)

摘 要 提取文献中高频出现的关键词,通过倒排索引的方法将关键词在摘要中定位,挖掘出摘要中隐含的与关键词能构成固定搭配的语义词组,并运用文本计量的方法追踪词组近年来的动态变化。利用关联影响度矩阵对语义词组进行了网络分析。实验结果表明,文献摘要中隐含的知识片段更能反映学科的发展趋势。

关键词 倒排索引,文本计量,关联影响度矩阵,社会网络分析
中图法分类号 TP113 **文献标识码** A

Novel Mining Implicit Text Fragment from Abstract Scheme

DAI Lu DING Li-xin XUE Bing
(State Key Lab of Software Engineering, Wuhan University, Wuhan 430072, China)

Abstract This paper extracted high-frequency keywords appearing in the literature, then positioned the abstract through inverted index, mined the fixed semantic phrases with keywords in the abstract, and tracked the dynamic changes phrases in recent years by text bibliometric. By using the related affect matrix to establish associated network, the association between the semantic phrases was analysed and figured out. The experimental results show that the literature summary implicit knowledge fragments can better reflect the trends of disciplines.

Keywords Inverted index, Text bibliometric, Related affect matrix, Social network analysis

1 引言

随着科学技术的进步,各种学科的文献逐步增多,读者在检索需求文献的时候往往是从文章的题名、摘要以及关键词入手。检索者检索论文首先看到的是摘要部分,它是科技论文的重要组成部分,以提供文献内容梗概为目的,不加评述和补充解释,简明、确切地记述文献重要内容的短文[1],其作用是使读者不用阅读论文全文即能获得必要的信息,同时为科技情报和检索提供方便[2]。关键词是从摘要中提取出来的能标示文献关键主题内容的自然词。关键词的运用,主要是为了适应计算机检索的需要[2]。通常情况下,如果某个领域成为热点,关于该领域里的一些科学研究方法的具体名称就会经常出现在关键词中,用另一种说法就是在某个阶段一些研究方法的具体名称出现在关键词中的频率比较高的时候标示着这些方法正在被大多数人所关注。这些高频率出现的关键词组为读者了解学科动态追踪学科热点提供了很大的帮助。然而事实上,我们在阅读文献时发现文章中信息量越充足的摘要越能反映学科发展的概况。比如在我们对学科进行探索发展的同时,该学科的固有名称也从最原始的词干慢慢演变,比如最典型的 data mining,随着数据挖掘技术的成熟演变为成现在的 Web data mining 或者 text data mining 等不同分支。这些标示着新动态的主题词组往往会最先出现在文章的摘要中,而作者很可能在关键词里将其忽略。因此,本文将通过关键词在摘要中的定位找出新的固定搭配词组,并得出了摘要中这些隐含的知识片段有可能成为讨论热点被大家所关注的事实。

2 挖掘摘要中固定搭配语义词组

在基于 SCI 收录的数据库中,随机抽取了 8479 篇科技论文作为实验材料,用文本抽取模式去除噪声,共提取了 3684 个关键词,并从这些关键词中提取前 252 个经常出现的关键词,称这前 252 个经常出现的关键词为高频词。

2.1 建立倒排索引

1. 读入源数据文档(data mining 10000. txt,该文档由大量文献记录组成,每条文献记录由 UT-Unique Article Identifier 字段进行标识)。

2. 将根据 UT-Unique Article Identifier 字段标识的一条条文献记录切分出来,然后对每条文献记录的所有属性进行分离(作者、主题、关键字和摘要等),最后把这些信息通过 Lucene 开发包组织起来,建立倒排索引。

3. 建立倒排索引[3]使用的是 Lucene 开发包里的标准分析器(StandardAnalyzer)。标准分析器(StandardAnalyzer)会把读入的字符串内容中所有的空格、停用词(停用词即在文章中常见的、对确定文章的主题毫无意义的一些词,如 a, the

到稿日期:2012-09-26 返修日期:2012-12-12 本文受国家自然科学基金(60975050,60903168)资助。
戴 璐(1987—),女,博士生,主要研究方向为数据挖掘,E-mail:qjdxyxx@yahoo.cn;丁立新(1967—),男,教授,博士生导师,主要研究方向为智能计算;薛 兵(1989—),男,硕士生,主要研究方向为演化计算。

等)、标点符号去掉,经过前述预处理之后,将字符串中剩下的每个单词按照单词——文献记录号这种结构建立索引。另外需要说明的是,Lucene 的索引结构既保存了正向信息,也保存了反向信息。

4. 最终,程序会在一个指定的目录下生成索引文件。

2.2 同义词表的建立

1. 读入源数据文档中每条文献的所有关键词。

2. 创建一个 map 容器用于存放同义词表,键为一系列关键词的主干词,值为单词的列表[4]。关键词可能由一个单词或多个单词组成,如果关键词是一个单词,则该单词的原型词(所谓单词的原型词即单词的最简形式,比如一个名词,它可能加后缀 s 等等变成另外一个词,则该单词的所有变换形式的原型词都为该名词)为主干词;若关键词为多个单词,则将各个单词分离开,每个单词的原型都是主干词[5]。对于分离开的不能作为主干词的字符串,直接过滤掉,比如纯数字、单字母等。

3. 每读入一个关键词,提取它的所有主干词,对于每一个主干词,若 map 容器中的键值中存在它的原型词,则加入该键对应的值列表中;若 map 容器中的键值中不存在它的原型词,则创建以该主干词为键的、值列表中只有该关键词的键值对。

2.3 关键词定位表的建立

1. 上面已经得到了同义词表,遍历该同义词表,循环遍历 map 的键的单词,对于每一个主干词,取出它对应的关键词,关键词可能只有一个单词,也可以有多个单词。

2. 若关键词为多个单词,使用 Lucene 创建一个多关键字查询类(PhraseQuery);若关键词为一个单词,使用 Lucene 创建一个词条类(TermQuery)。

3. 在已索引的文献记录的摘要中进行查询,若一条文献记录的摘要中包含有该关键词,则将该关键词在摘要中的所有位置都找出来,并加入一个列表中。

4. 输出关键词在该文献记录摘要中的位置信息索引,其格式为 word:〈doc,freqn[index1,index2,…,indexn]〉,word 为关键词,doc 表示该文献记录在源数据文件中是第几条文献记录,freqn 为关键词在文献记录摘要中出现的次数,方括号里面的 index1,index2,…,indexn 分别为关键词出现在摘要中的位置。

5. 处理完一条文献记录之后,循环遍历后一条文献记录,将处理的结果添加到前一个结果后面,直到处理完所有文献记录,统计关键词在所有文献记录摘要中出现的频率。因此,每一个关键词最终的位置信息索引格式为:

word:〈doc1,freq1n[index11,index12,…,index1n]〉

〈doc2,freq2n[index21,index22,…,index2n]〉

…

〈docn,freqnn[indexn1,indexn2,…,indexnn]〉

TOTAL:num

num 为关键词在所有文献记录摘要中出现的频率。

6. 同义词表遍历完之后,输出所有关键词在摘要中的位置结果。

在位置信息索引中,对每个词项,以如下方式存储倒排记录:文档 ID:〈位置 1,位置 2,…〉,每个记录中同时也保存了文档中的词项频率信息[6]。位置信息索引可用于短语查询,K 词近邻搜索。为处理短语查询,仍然需要访问各个词项的倒排记录表。这里可以采用最小文档频率优先的策略,从而可以限制后续合并的候选词项的数目。采用位置索引会大大增加倒排记录表的存储空间,大概是非位置索引大小的2~4倍,而压缩后的位置索引大概为原始未压缩文档文本(去除)标记信息的 1/3~1/2[7]。

2.4 找关键词与其前面或后面的词组合为有意义的词

1. 载入关键词定位表。

2. 对于同一组同义词(主干词相同的关键词为同义词),遍历其中的每一个关键词的位置信息。对于关键词,先定位它在文献摘要记录中的位置,然后将该关键词与它前面的 n 个词进行组合($n=1,2,3,\cdots$),若它与 n(n 为最大的值)个词进行组合有意义,则说明找到了一个结果。若在该组同义词内,之前出现过该组合词,则将该组合词的频率加 1;否则把它的频率置为 1。

3. 将关键词与它前面的词进行组合之后,再将该关键词与它后面的 n 个词进行组合($n=1,2,3,\cdots$),若它与 n(n 为最大的值)个词进行组合有意义,则说明找到了一个结果。若在该组同义词内,之前出现过该组合词,则将该组合词的频率加 1;否则把它的频率置为 1。

4. 一组同义词处理完后,再换另一组同义词进行处理,直到全部处理完,输出结果(同义词的主干词,前组合词及对应的频率,后组合词及对应的频率)。否则转到 2。

高词频出现的次数相对较高但并不意味着每一个高频词都能在摘要中找到固定搭配,比如关键词 knowledge extraction 出现了 741 次,但该关键字在摘要中却没有某个具有实际语义的单词经常和它一起出现,形成固定搭配,成为固定搭配语义词组。但一些高频词在文章摘要里出现的时候,会有一些有实际语义的词与之搭配,它们形成某种意义上的固定搭配,例如 Functional genomics 和 decision trees,当它在关键词中出现时,经常会在摘要中出现 functional genomics approaches 和 fuzzy decision trees。像这样在论文的关键词中出现的词或者词组称为词干,这些词或者词组与文章摘要中有实际语义的词形成的一定的固定搭配,成为一个具有新语义的词组,该词组称为固定搭配语义词组(简称为固配语义词)[8]。

当这种固定搭配语义词组出现的频率比较高,占词干的比例比较大的时候,代表着该词组所代表的知识点正在被关注,或者该知识方向已经趋向于成熟,被广泛应用。如果该固定搭配语义词组占的比例比较小,说明该单词或者词组所在的研究领域或者是研究方向并不成熟,或许以后会成为一个新的研究热点[9]。

3 实验结果和分析

在基于 252 个高频词的 29 个词干所形成的 30 个有具体

语义的固定搭配语义词组中(见表 1),以典型的词组为例运用传统的文献计量的分析方法[10],即统计该固定搭配语义词组在摘要中出现次数的浮动来探讨该知识点近年来的变化趋势。

表 1　摘要中找到的隐含的固定搭配语义词组以及在摘要和关键词中出现的情况对比

在关键字中出现的形式	在关键字中出现的次数	在摘要中出现的形式
Decision trees	2940	fuzzy decision trees
pattern recognition	2447	structural pattern recognition
subspace clustering	169	Subspace Clustering based on Information
genetic programming	895	grammar based genetic programming
genetic programming	895	cellular genetic programming
scientific discovery	118	scientific discovery systems focus
multidimensional indexing	4	multidimensional indexing structures
binary data mining	33	real-life binary data mining
logic data mining	29	fuzzy logic data mining
Web information systems	21	Web information systems automatically
sequential pattern	896	sequential pattern mining algorithm
machine learning	8520	proper machine learning algorithm
multiple sequence alignment	182	including multiple sequence alignment
workflow management	506	contemporary workflow management
structural genomics	353	northeast structural genomics
multimedia data mining	313	proposed multimedia data mining
multimedia data mining	313	subspace-based multimedia data mining
multimedia data mining	313	multimedia data mining framework
Web mining	951	abstract Web mining
data mining	1663	fuzzy data mining
data mining	1663	mobile data mining
privacy preserving data mining	543	privacy preserving data mining techniques
grid computing	416	grid computing service technology
information fusion	361	Web information fusion
information retrieval	1376	support effective multimedia information retrieval
dimensionality reduction	780	multifactor dimensionality reduction
Functional genomics	494	functional genomics approaches
pattern recognition	2447	statistical pattern recognition
machine learning	8520	genetic algorithm based machine learning
frequent itemset	529	maximal frequent itemsets mining

图 1 是 30 个有具体语义的固定搭配语义词组占词干的百分比图。从图 1 中可以看出有的固定搭配语义词组占词干的百分比非常高,达到 57.96%,说明它在众多研究领域中是热点,经常被关注,甚至说明该知识点已经成熟或者趋于成熟,被广泛应用。有的固定搭配语义词组占词干的百分比非常低,达到 0.04%,说明它在众多研究领域中逐渐不被关注,或者被新的知识和技术所取代。

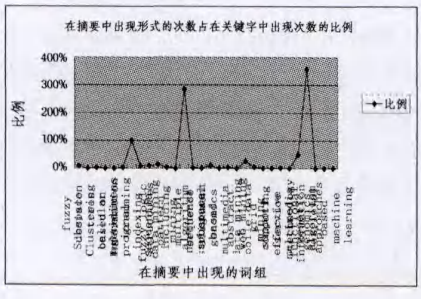

图 1　词干在摘要和关键词中出现次数的对比

如图 2 所示,该词组 1996－2000 年在关键词中出现了 53 次,在 2001 年－2005 年出现了 68 次,在 2006－2009 年出现了 397 次。图 3 中,该词组在 1996－2000 年出现了 57 次,在 2001－2005 年出现了 201 次,在 2006－2009 年出现了 178 次。这两个图都表示该固定搭配语义词组代表的知识点在某段时间不被关注,后来逐渐被关注,被应用,成为热点。

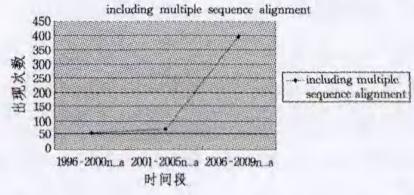

图 2　including multiple sequence alignment1996－2009 年的变化趋势

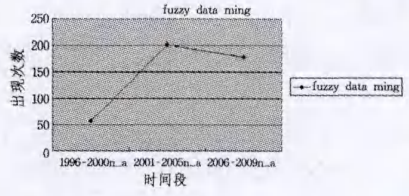

图 3　fuzzy data mining1996－2009 年的变化趋势

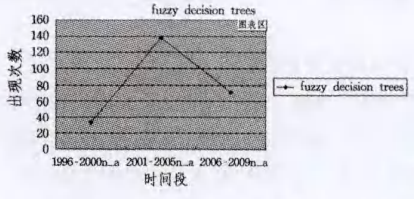

图 4　fuzzy decision trees 在 1996－2009 的变化趋势

从图 4 中可以看出,该词组在 1996－2000 年出现的次数较少,在 2001－2005 年出现的次数较多,在 2006－2009 年出现的次数又逐渐变少。说明该固定搭配语义词组开始不被关注,后来逐渐被关注,成为热点,并且逐渐成熟,后来又逐渐不被关注或者逐渐被新的知识和技术所取代。

4　固定搭配词组的网络分析

4.1　关系影响度矩阵的建立

从源数据文件的摘要中选定有语义的词组,从这些有语义词组中找出这样的关系:哪些词组和另外一些词组联系紧密些,或者说有哪些词组出现在某篇摘要中时,另外一些词组也很有可能会出现。如果词组 a 在文章(文章号假定为 k)摘要中出现了 m 次,词组 b 在文章(文章号假定为 k)摘要中出现了 n 次,那么词组 a 与词组 b 的联系基数为 $m*n$,并画出它们之间的关系图。关系图可能还隐含了其它的关系,比如同义词之间的联系等。根据要分析的关系,构造相应的矩阵,再由矩阵导出相应的关系图[10]。我们首先给出关联度矩阵的定义。

定义 1　给定关联影响信息系统 $S=(U,A,V,F)$,$A=C\cup D$,矩阵 $M=|m_{ij}|$,$m_{ij}=\mathrm{Inter}(m_i->m_j)$,$m_i,m_j\in A$。$M$ 称为关联影响度矩阵[11],见表 2。

表 2　关联影响度矩阵

	F1	F2	…	Fm	D1	…	Di
F1	1	Inter(F1->F2)	…	Inter(F1->Fm)	Inter(F1->D1)	…	Inter(F1->Di)
F2	Inter(F2->F1)	1	…	Inter(F2->Fm)	Inter(F2->D1)	…	Inter(F2->Di)
F3	Inter(F3->F1)	Inter(F3->F2)	…	…	Inter(F3->D1)	…	Inter(F3->Di)
…	…	…	…	…	…	…	…
Fm	Inter(Fm->F1)	Inter(Fm->F2)	…	1	Inter(Fm->D1)	…	Inter(F2->D1)
D1	Inter(D1->F1)	Inter(D1->F2)	…	Inter(D1->Fm)	1	…	Inter(D1->Di)
…	…	…	…	…	…	…	…
Di	Inter(Di->F1)	Inter(Di->F2)	…	Inter(Di->Fm)	Inter(Di->D1)	…	1

通过关联影响度矩阵，构造这些有语义的词组之间的关联分析，从而找出某些有语义的词组之间的联系。

下面给出该关系图的构造过程：

1. 根据找出的 n 个有语义的词组，构造一个 $m * n$ 的矩阵 A，m 为所有的文献记录条数，n 为有语义词组的个数，最初矩阵所有元素的值都为 0，即从 m 篇摘要中找出 n 个有语义的词组，构造一个所有元素都为 0 的 $m \times n$ 的矩阵 A。

2. 从第 i 篇摘要开始，统计在这 n 个有语义词组中每个词组在该摘要中是否出现，如果有语义的词组在该摘要中出现，则将该条文献记录和该词组对应的交叉点的数据改为 1，否则为 0。从该摘要最前面开始遍历，从第一个有语义的词组开始，若某个有语义的词组出现，就将该条文献记录和该词组对应的交叉点的数据改为 1，表明该词组在该摘要中出现，并结束对该词组的遍历，开始下一个词组的查找；否则，就将该条文献记录和该词组对应的交叉点的数据改为 0，表明该有语义词组没有在该摘要中出现。对该篇摘要遍历 n 遍，得出这 n 个有语义词组是否在该摘要中出现。同时也得出矩阵 A 的第 i 行元素。

3. 当 $i \leqslant m$ 时，转到 2；否则转到 4。这时循环遍历了所有文献记录，对于有语义词组是否在某条文献记录中出现，在该条文献记录和有语义词组对应的交叉点处已作修改，如果出现，该处的值为 1；如果没有出现，该处的值为 0。第 m 篇摘要遍历结束后，得到文献记录与有语义词组的矩阵 $Am \times n$。即循环遍历 m 篇文献记录，得出 n 个有语义词组的每个有语义词组在 m 篇文献中是否出现，并记录在矩阵 $Am \times n$ 相应的位置。

4. 根据矩阵 $Am \times n$ 可通过数学运算得到一个 $n * n$ 的矩阵 B。即将矩阵 $Am \times n$ 转置，得到转置矩阵 AT，根据 $AT * A$ 可得到矩阵 $Bn \times n$，n 为有语义词组的个数。矩阵 B 为词组与词组的矩阵，矩阵 B 某行某列交叉的值，即词组 a 与词组 b 的交叉点的值是词组 a 与词组 b 在源数据文件中所有文献记录中同时出现的文献次数。

5. 将矩阵 B 导出，用一个 excel 表进行存储，第一行和第一列都为词组，并将 excel 表中的矩阵 B 复制到 UCINET 软件中的 Matrix spreadsheet editor 中。

6. 在 UCINET 软件的 Matrix spreadsheet editor 界面上，将矩阵 B 保存为后缀为 XXX. # # h 的文件（XXX 可由用户自己任意命名）。

7. 关闭 Matrix spreadsheet editor，点击 UCINET 软件上的 NetDraw 按钮，在出现的界面上导入 UCINET network dataset（即第 6 步导出的名字为 XXX. # # h 的文件）之后，软件会生成一个图，然后根据 NetDraw 菜单上的按钮调整图形（线和结点颜色及大小、整个图的布局等等）。

4.2　社会网络分析构建的固定搭配语义词组

本文随机选出了 25 个有语义的词组，构造出了关联影响度矩阵，并形成了社会网络图谱，从而得到语义词组之间的联系。这些有语义的词组之间有线连接，而且有的连接线粗，有的连接线细，一根线连起来的两个有语义的词组，说明它们在同一篇摘要中出现过，而且两个有语义的词组在多篇摘要中同时出现的次数越多，它们之间的连线就越粗，例如有语义的词组 Subspace Clustering based on information 和 nonlinear mapping network 之间的连线比较粗，说明它们在多篇摘要中同时出现，即当 Subspace Clustering based on information 在摘要中出现的时候，nonlinear mapping network 也在该篇摘要中出现，或者当 nonlinear mapping network 在摘要中出现的时候，Subspace Clustering based on information 也在该篇摘要中出现，属于强关联[12]。fuzzy decision trees 和 cellular genetic programming 之间的连线比较细，说明它们在同一篇摘要中出现的次数比较少，即 fuzzy decision trees 在某篇摘要中出现的时候，cellular genetic programming 在该摘要中也出现过，但出现的次数很少，词组 fuzzy decision trees 和 cellular genetic programming 的联系相对不紧密。

结束语　文献中的关键词是用来方便读者检索文献，因此关键词在一段时间内出现的频率从某种角度上能反映学科的动态。实验表明，作者在提出关键词的时候往往会忽略摘要中和关键词搭配的语义词组，这些高频出现的固定词组很有可能会成为被读者研究的另一个知识热点。

参 考 文 献

[1] 袁鼎荣，钟宁，张师超. 文本信息处理研究述评[J]. 计算机科学，2011，38(2)：9-13

[2] 一种基于主题的概率文档相关模型[J]. 计算机科学，2008，35(10)：178-180

[3] Manning C D，Raghavan P，Schütze H. 信息检索导论[M]. 王斌，译. 北京：人们邮电出版社，2010：29-30

[4] 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

[5] 孟祥福，张霄雁，马宗民，等. 基于语义相似度的 Web 数据库不精确查询方法[J]. 计算机科学，2012，39(4)：154-158

[6] 邱均平，王菲菲. 基于文献计量的国内外社会网络分析研究比较[J]. 情报资料工作，2011(1)：36

[7] 季淑娟，董月玲，王晓丽. 基于文献计量方法的学科评价研究[J]. 情报理论与实践，2011(11)：27-33

[8] 陈君，唐雁. 基于 Web 社会网络的个性化 Web 信息推荐模型[J]. 计算机科学，2006，33(4)：185-187

[9] 马君华. 粗糙集属性约简与聚类算法及其在电力自动化中的应用与研[D]. 武汉：华中科技大学，2010

[10] 李杰，徐勇，王云峰，等. 面向个性化推荐的强关联规则挖掘[J]. 系统工程理论与实践，2009，29(8)：72-76