基于词汇链的关键短语抽取方法的研究

刘 铭 王晓龙 刘远超
(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

摘 要 文中提出一种基于词汇链的关键短语抽取算法,算法首先通过构造多条词汇链来表达文章的多条叙事线索,并从多条词汇链中抽取富含主题信息的强链代表文章着重叙述的信息,然后从强链中选取能够从不同侧面充分表达强链所述信息的短语作为文章的关键短语.实验表明该算法抽取的关键短语能够更全面地覆盖文章的主题信息.算法消除了多个关键短语表达同一主题信息的冗余性,同时可以根据文章主题的分布动态确定输出的关键短语的数量,其效果明显优于采用统计信息进行关键词抽取的方法.

关键词 词汇链;知网;中心词聚类;关键短语;词义获取
中图法分类号 TP391 **DOI号**: 10.3724/SP.J.1016.2010.01246

Research of Key-Phrase Extraction Based on Lexical Chain

LIU Ming WANG Xiao-Long LIU Yuan-Chao
(*School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001)

Abstract A novel algorithm for key-phrase extraction based on lexical chain is proposed in this paper. By constructing lexical chains for each article, the article's multiple depiction clews can be reflected, and some strong lexical chains with high quality can be extracted to represent main content of this article. After previous operations, key-phrases, which can fully express topic information of strong chain from different aspects, are extracted. Experiments demonstrate that key-phrases from this algorithm can cover article's topic more completely. This algorithm can remove redundancy that different key-phrases reflect same meanings, and can dynamically decide the size of output key-phrase set by distribution of topic information. This method outperforms the method which uses statistics to perform extraction.

Keywords lexical chain; HowNet; central-word clustering; key-phrase; acquisition of meaning of word

1 引 言

随着网络的普及,人们每天接触的信息与日俱增,如何快速并准确地掌握大量信息所描述的内容在人们的日常生活中变得越来越重要.关键词标注技术是上述问题的一个很好的解决办法,好的关键词能够使读者快速掌握文章的主要内容,加深读者对文章的理解.关键词抽取一直是文本挖掘领域的主要研究问题,同时该技术还可以应用于其它领域,例如大量的图书馆系统和信息检索系统使用关键词抽取技术构造文件索引[1-2];许多文本挖掘系统以关键词所在的句子作为文摘句[3-4];很多聚类和分类算法也使用关键词算法构造文章的特征向量以提高算

收稿日期:2008-09-05;最终修改稿收到日期:2010-04-21.本课题得到国家自然科学基金重点项目(60435020)、国家"八六三"高技术研究发展计划目标导向类课题(2006AA01Z197,2007AA01Z172)资助.**刘 铭**,男,1981年生,博士研究生,研究方向为聚类分析、文本挖掘.E-mail: mliu@insun.hit.edu.cn. **王晓龙**,男,1955年生,教授,博士生导师,研究领域为信息检索、文本挖掘. **刘远超**,男,1971年生,副教授,研究方向为聚类分析、人工智能.

法的准确度同时降低特征空间的维度[5-6].

目前多数关键词抽取算法是利用词的统计信息判断词的重要性[1-3],并选取超过一定阈值的词作为文章的关键词.基于这种方法提出了多个关键词衡量函数,包括 TF/IDF[7]、熵函数[8]、分布系数[9]等.许多机器学习算法也应用于关键词抽取中,例如朴素贝叶斯算法[10]、C4.5[11]、决策树[12]和最大熵算法[13].上述算法通过训练语料获得抽取函数,然后选取能够使抽取函数得到最大值的词作为关键词.然而由于中文文档包含信息的多样性,使得现实应用中很难获得一个通用的抽取函数或模型用于关键词抽取.也有算法考虑了相似词在文中的分布情况,抽取具有大量相似含义词的特征词作为关键词[14-15],其有结合统计的方法,也有结合词典的方法.结合词典的关键词抽取方法多以 HOWNET 和 WORDNET 作为计算词语相似度的基准词典,其中 HOWNET 多用于中文文本关键词的抽取,刘群、李素建等即通过 HOWNET 计算词语间相似度,然后通过聚类方式获得词类,并选择最能反映文档主题信息的词类抽取关键词[16],该方法抽取的关键词能够在一定程度上防止信息冗余,但是大量的无关词语降低了关键词抽取的准确性.词典 WORDNET 多应用于英文领域中,由于 WORDNET 是以词类组织词语的,因此使用该词典能够直接完成词类的划分[17],但是该方法存在两个比较显著的问题,一是现实应用中的词语大多是一词多义的,其应该被划分到多个类别中,而 WORDNET 显然没有考虑到这个问题;二是该词典没有考虑到词语之间的相关性,即词类的划分仅仅是以词语之间的相似性来度量的.统计的方法也广泛应用于关键词抽取中横量词语之间的相互关系[18-19],但是统计方法计算量过大,并且需要大量的统计语料.

瑞典斯德哥尔摩大学的博士 Anette 的论文[20-21]对关键词抽取做了较为深入的研究,其论文的主要思想是:首先获得词在文中的词性、出现的位置、词频(TF)、文档倒排频率(DF)等统计信息,然后构造统计模型预测这些统计量的重要性并进行融合,选取融合后得分高的多个词作为关键词予以输出.可以看出上述方法是通过融合文中词的统计信息来确定词的重要性,但是值得注意的是某些具有高统计信息量的词并不一定能够确切反映文章的主题,同时单个词富含的信息量较少,反映的信息也不够清晰.本文即针对上述问题提出一种关键短语抽取算法,算法首先通过构造词汇链对文章主题进行分析,分析文中包含的多条主题线索,在此基础上选取能够充分代表这些主题线索且富含更多信息的短语作为关键短语,使得生成的关键短语能够确切反映文章叙述的主题信息.

2　词义获取

本文以词汇链反映文章的主题信息,词汇链是 1991 年由 Hirst 首先提出的,以相关或相似的词语构成的一条链.词汇链与文本的结构有一种对应关系,它提供了关于文本结构和主题的重要线索[22].词汇链是由围绕文中某主题的许多相关词组成的集合体,因此在构建词汇链时需要知道词在某个上下文中的确切含义.本文以“知网”(HOWNET)作为词义获取的语义词典.“知网”是由董振东博士完成的中文语义辞典,其定义了 1500 多个义原,并通过义原反映中文词义[23].“知网”以 DEF 表达词条的语义信息,由两部分组成,分别为基本义原和关系义原.其中基本义原能够在很大程度上反映 DEF 的含义,关系义原代表了 DEF 的关系结构特性.“知网”以树形组织义原,越相似的义原在义原树内的位置越接近.

目前基于“知网”的词义确定多是将多义词的词义定义到该词对应的 DEF 集合中唯一的一个 DEF 上[24].但是观察“知网”可知,“知网”中对于 DEF 的定义过于严格,同一词的多个 DEF 在现实中的区别并不严格,并且 DEF 中的基本义原在很大程度上决定了 DEF 的含义,至少对于本文的应用来说,这个结论是成立的.因此本文将“知网”中基本义原相同的一个 DEF 集合视为词条的一个义类,而本文中对于词义的获取就是对一个含有多个义类的多义词找到其在某个确定的上下文中对应的那个义类.

词的上下文语境在很大程度上决定了多义词在此上下文中的词义,但是只有少量的上下文信息对多义词的词义具有决定性的影响[25],因此本文以待确定词义的多义词的前两词和后两词作为词的上下文信息,并依此获取多义词在文中对应的义类.

如图 1 所示,假设文档 Doc 中的词序列为 M_1、M_2、M、M_3、M_4,其中 M_1、M_2、M_3、M_4 为 M 的上下文信息.图中的顶点(圆)代表每个词对应的义类,顶点间的边为义类间的关联度.为了清晰地显示图中每个顶点的含义,我们可将图 1 看作为一个矩阵,其行对应的是词序列,包括待确定词义的词 M 和 M 的上下文词 M_1、M_2、M_3、M_4.列对应是词语包含的

多个义类，即多义词的不同词义. 从图中可见，词 M_1 对应的列包含两个顶点，即该多义词有两个词义，词 M_2 对应的列包含一个顶点，即该词为单义词. M 在此上下文中的词义即是在图 1 中寻找一条从 M_1 开始到 M_4 结束的连通分量，该连通分量的边的权值之和最大，然后以此最大连通分量经过的 M 的那个义类作为 M 在此上下文中的词义. 其最大连通分量可以采用 Dijkstra 算法进行寻找，在此就不赘述了[26]. 按上述方法依次处理文档 Doc 中的所有词即可获得文档词空间中所有词的词义. 式(1)即为边权值的计算函数，其描述了两个义类联系的紧密程度.

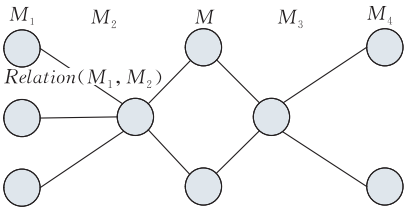

图 1 义类关系图

上文介绍的词义确定依据以下朴素的想法：对于多义词的每个义类，某个义类与该词的某个上下文的关系越紧密，则此义类作为该词在此上下文中的词义就越有可能.

$$R(M_{im}, M_{jn}) = \frac{SW(M_{im}, M_{jn}) + CW(M_{im}, M_{jn})}{2} \tag{1}$$

式(1)中 M_{im} 代表词 M_i 的第 m 个义类，此关联度公式为义类间的相似度与相关度的平均值. 其中相似度代表了两个义类描述信息的相似性，可以由两个义类的基本义原在义原树内的位置和两个义类是否具有相同的结构来决定. 相关度代表了两个义类描述的信息之间是否相互关联，主要由两个义类的 DEF 结构的交叉性来决定.

$$SW(M_{im}, M_{jn}) = \lambda \times FS(M_{im}, M_{jn}) + (1-\lambda) \times RS(M_{im}, M_{jn}) \tag{2}$$

式(2)计算了两个义类间的相似度. 以加号为界，第 1 部分计算了两个义类中基本义原的相似度，其通过基本义原在义原树内的位置进行衡量；第 2 部分计算了两个义类中关系义原的相似度，其通过义类的关系结构的相似性进行衡量. 参数 λ 对应于两部分的重要性，由于基本义原较能反映词的主要信息，因此 λ 的设置偏重于第一部分，实验中设 λ 为 0.6.

$$FS(M_{im}, M_{jn}) = 1/(Position(M_{im}, M_{jn}) + 1) \tag{3}$$

式(3)中 $Position(M_{im}, M_{jn})$ 代表了两个义类 M_{im} 和 M_{jn} 的基本义原在义原树内的层次差，如果两个基本义原不在同一义原树内则该值为 ∞. 可以看出如果两个基本义原在义原树内的层次越接近，即两个基本义原越相似，则 $FS(M_{im}, M_{jn})$ 的值越大.

$$RS(M_{im}, M_{jn}) = \frac{IS(M_{im}, M_{jn})}{RC(M_{im}) + RC(M_{jn})} \tag{4}$$

式(4)中 $IS(M_{im}, M_{jn})$ 指两个义类 M_{im}、M_{jn} 的关系义原集合的交集大小，代表了两个义类的关系义原的相似程度. $RC(M_{im})$ 指义类 M_{im} 具有的关系义原总数.

式(5)计算了两个义类间的相关度，此相关性反映了两个义类所反映的信息之间是否相互关联，例如是否具有从属、支配等关系.

$$CW(M_{im}, M_{jn}) = \frac{I(M_{im}, M_{jn}) + I(M_{jn}, M_{im})}{RC(M_{im}) + RC(M_{jn})} \tag{5}$$

其中 $I(M_{im}, M_{jn})$ 指义类 M_{im} 的关系义原集合中是否包含 M_{jn} 的基本义原，如果包含，值为 1，否则为 0. 由于基本义原代表了义类的主要信息，关系义原代表了义类的关系特性，因此 $I(M_{im}, M_{jn})$ 能够说明 M_{jn} 反映的信息是否与 M_{im} 具有一定的关联关系[27].

3 词汇链构造

如文献[22]所述，词汇链是以文中相关或相似的词语组成的链状集合体，每条链能够代表文章所描述的某个子主题信息，构造的多条链能够反映文章的多条叙事线索. 本文即通过构造词汇链从各个侧面反映文章的主题信息，并从词汇链中抽取能够充分代表该链所述信息的短语作为关键短语.

按文中第 2 节所述方法进行词义获取后即可对文章构造全文词汇链，具体方法就是扫描文档 Doc 的词空间(WordSet)，然后选择与当前处理的词具有最大相似度的词汇链并将该词插入到此词汇链中，最后选择权重大于平均值的词汇链作为强链以反映文档 Doc 所描述的主题信息.

式(6)为词空间中的词 M_q 与词汇链集合中的链 L_p 的相似度计算公式.

$$SC(M_q, L_p) = \max_{t=1}^{|L_p|}[SW(R(M_q), R(LW_{pt}))] \tag{6}$$

其中 $|L_p|$ 为词汇链 L_p 包含的词数，LW_{pt} 为链 L_p 中的第 t 个词，$R(LW_{pt})$ 为 LW_{pt} 经文中第 2 节词义获取后对应的义类. 我们以词 M_q 与链 L_p 中所有词的最大相似度作为 M_q 与 L_p 的相似度[22].

词汇链构造中需要预先设定词与词汇链之间的

相似度阈值,而由式(6)可见,我们以词与词汇链包含的所有词的最大相似度作为词与词汇链的相似度,因此词与词汇链之间的相似度阈值也就是词与词之间是否相似的阈值.实验发现如果词与词之间的相似度超过 0.7,则两个词较为相似.例如文中第 6 节实验的第一部分中"种植"和"栽种"在文中的相似度为 0.78,而"种植"和"培育"在文中的相似度为 0.53.

式(7)为词汇链 L_p 的权值计算公式.

$$WC(L_p)=\sum_{t=1}^{|L_p|}W(LW_{pt})\times\log_2(|L_p|) \quad (7)$$

其中的词权重即 $W(LW_{pt})$ 是根据词 LW_{pt} 是否被标题包含、出现的段落位置、所在句中是否含有线索词、词分布等统计量进行加权回归后得到的权重[28].由式(7)可见,词汇链的权重与词汇链包含的词数和词权重有关,如果词汇链包含的词数越多说明该词汇链反映的信息在文中的分布越广,如果词汇链包含的词权重越大说明该词汇链反映的信息在文中越重要.

4　关键短语抽取

上文构造的多条强链可以反映文章的多条叙事线索,然而每条线索均有不同的侧重点.本文即通过抽取强链中代表不同侧重点的中心词来表达上述侧重信息.由于作者在叙述文章时习惯将近似的词语交互使用,因此本文以每条强链中的每个候选中心词作为聚类中心,然后在强链内选择与聚类中心的相似度大于 0.7(阈值设定原因如文中第 3 节所述)的词插入到作为聚类中心的候选中心词代表的词类中去,以获得与候选中心词相似的词语在词汇链内的分布情况.按式(8)计算每个候选中心词的权值,并选取大于平均权重的中心词代表文中多条叙事线索的不同侧重点,以这些中心词的并集作为文章的中心词集合.本文以文档词空间的平均词频作为阈值,以每条强链中满足上述阈值的词作为该强链的候选中心词.

$$W(CW)=\sum_{l=1}^{|WS(CW)|}W(F_l)\times\log_2(|WS(CW)|) \quad (8)$$

式(8)中 $|WS(CW)|$ 为候选中心词 CW 代表的词类 $WS(CW)$ 所包含的词数,$W(F_l)$ 为 $WS(CW)$ 包含的第 l 个词的权值,具体计算方法可参见式(7).可以看出每个候选中心词的权值由以此候选中心词代表的词类所包含的词权重以及该词类所包含的词数共同决定.这样即可通过分析与候选中心词 CW 具有相似信息的词在词汇链内的分布情况来判断此候选中心词所反映的信息在文中的重要性.

短语要比词含有更丰富的信息,可读性更强,因此本文期望以短语来覆盖更多的主题信息.现实中的短语大多以二元和三元结构居多,则本文对于短语的构建也是基于二元和三元短语结构.本文采用文献[29]中介绍的词性构成规则作为短语搭配模板,对满足词性搭配模板的短语统计短语内词的同现率,如式(9)、(10)所示,并截取超过一定比率的短语作为关键短语.同现率能够反映两个或多个词之间是否具有相关性,两个或多个词的同现率较高,说明这两个或多个词经常一起出现,具有很强的相关性,作为短语的可能性很大[30].

$$FreCoOccur(w_1,w_2)=\frac{P(w_1,w_2)}{P(w_1)\times P(w_2)} \quad (9)$$

$$FreCoOccur(w_1,w_2,w_3)=\frac{P(w_1,w_2,w_3)}{P(w_1)\times P(w_2)\times P(w_3)} \quad (10)$$

$P(w_1,w_2)$ 指 w_1 和 w_2 两词在语料中满足语法构成规则作为短语出现的次数,$P(w_1)$ 指词 w_1 在语料中出现的次数.本文以 1998 年的人民日报作为同现率的统计语料.

下面介绍关键短语的抽取步骤:

1. 初始化.设 *OGS*(Output Gram Set)为关键短语输出集合,*OGST*(Output Gram Set Temp)为候选短语集合,设上文产生的中心词集合为 *CWSet*;

2. 短语选取.按照词性模板对文章进行筛选,选出满足条件的二元、三元短语,从超过平均频度的短语中抽取同现率超过 90%的短语压入到 *OGST* 中;

3. 去重.删除 *OGST* 中被三元短语包含的二元短语;

4. 筛选.从 *CWSet* 中删除被 *OGST* 中的短语包含的中心词,同时将包含该中心词的短语压入到 *OGS* 中.将 *CWSet* 中没有被任何短语包含的中心词也压入到 *OGS* 中;

5. 排序输出.计算 *OGS* 中短语的权重,短语权重为短语包含的词的权重之和.对 *OGS* 中的短语和中心词按其权重进行排序并输出,如果对输出的关键短语有数目上的限制则截断输出.

5　时间复杂度分析

如上文所述,本文介绍的关键短语抽取算法主要分为 3 个部分:词义获取、词汇链构造、关键短语抽取.词义获取部分的时间复杂度:算法需要顺序扫描文档的词空间以获得词义,同时算法在进行词义获取时要依次对待确定词义的词的多个义类进行处

理.假设分词及停用词过滤后文档的词空间的维数为 n,并设词在知网中具有的义类数最多为 k,则上述词义获取对应的连通图顶点数为 kn.当我们使用 Dijkstra 算法求解最短路径以获取词义时,其时间复杂度为 $O(k^2n^2)$.观察知网得知,词在知网中的最大义类数不超过 6,因此上述词义获取的时间复杂度即为 $O(n^2)$.

词汇链构造部分时间复杂度:算法需要顺序扫描文档的词空间以线性构造词汇链,同时在构造词汇链时,需要计算词空间中的每个词与每条词汇链的相似度.同样假设分词及停用词过滤后文档的词空间的维数为 n,则可知文档至多包含 n 条词汇链,即每条词汇链仅包含一个词,因此上述词汇链构造的时间复杂度最多为 $O(n^2)$.在词汇链构造算法中还需要计算每条词汇链的权重以选择词汇链集合中能够表现文档主题信息的强链,即强链选择的时间复杂度最多为 $O(n)$.则词汇链构造部分的总的时间复杂度为 $O(n^2+n)=O(n^2)$.

关键短语抽取部分时间复杂度:关键词抽取部分首先需要构造候选中心词并进行候选中心词聚类.在选择候选中心词时,需要计算强链中包含的每个词的权重以选择候选中心词,由此可知上述候选中心词选择时最多需要对文档词空间内的所有词计算权重,即文档的每条词汇链均为强链,则候选中心词选择的时间复杂度为 $O(n)$.在进行候选中心词聚类时,需要对每个候选中心词在强链内寻找与其相似的词语,假设文档词空间的维数为 n,那么候选中心词的数目最多为 n,同样需要与每个候选中心词计算相似度的词的数目最多为 n,因此候选中心词聚类的时间复杂度最多为 $O(n^2)$.在获得候选中心词类后即对每个词类计算权重然后选择中心词以生成中心词集合,易知候选中心词类的数目最多为文档词空间包含的词数 n,即每个词类仅包含一个词,则中心词集合生成的时间复杂度最多为 $O(n)$.在获得中心词集合后,即可进行关键短语抽取,其首先需要扫描文档词空间以生成候选短语,此部分的时间复杂度为 $O(n)$,然后扫描中心词集合以过滤掉被短语包含的中心词,并将不被短语包含的中心词予以输出,可以看出此部分最多需要扫描的词数为文档词空间包含的词数,即此部分的时间复杂度最多为 $O(n)$.综上所述,关键短语抽取部分的时间复杂度为 $O(n^2+4n)=O(n^2)$.

将上述 3 部分的时间复杂度进行叠加即可得到关键短语抽取算法的总的时间复杂度为 $O(n^2+n^2+n^2)=O(n^2)$.现今广泛使用的关键词抽取算法可以分为两类,一类是基于机器学习方法的,一类是基于词权重的.基于机器学习的关键词抽取算法大多通过训练语料获得一个抽取函数,然后通过抽取函数判断关键词,上述算法均可分为训练和抽取两个过程,其抽取过程非常快,大多是 $O(n)$,而训练过程的时间复杂度却极高,大部分都超过 $O(n^2)$.基于词权重的关键词抽取算法大多通过词的位置、频度、词类等统计信息计算词权来选择高权值的词作为关键词,此类算法不需要训练过程,但是需要扫描词空间以获得词空间内的每个词的分布情况来计算权重,即时间复杂度为 $O(n^2)$.比较算法的时间复杂度可知,本文所提算法的时间复杂度为 $O(n^2)$,即算法的时间复杂度与目前广泛使用的关键词抽取算法的时间复杂度相当.

6 实验结果及分析

关键词抽取技术大多应用于其它算法的预处理阶段,同时对关键词抽取结果的判定的主观性较大,即使对同一篇文档不同的人也会获得不同的关键词抽取结果,因此现实应用中很难找到标准的关键词抽取评测语料.本文以“任常霞先进事迹”、“印尼海啸灾难”、“圆明园水渗漏的治理”、“山野菜的种植”、“足球机器人比赛”为主题并采用搜索引擎 Google 进行检索,将检索得到的前 50 篇共 250 篇文档作为测试语料,分别为每篇文档手工标定 20 个关键短语.

首先我们结合一篇文章来分析词汇链构造以及关键短语抽取的结果,并将其与按照统计信息抽取的关键词进行对比.该文题目为《棚栽山野菜半亩收万元》①,主要介绍了“暖棚山野菜的种植带来巨大收益”.

对文章构造词汇链后输出权重位于前列的部分词链:

(1)山野菜;蔬菜;黄瓜;西红柿;芹菜;

(2)采摘;收获;

(3)经济效益;效益;

(4)种植;移栽;栽种;

(5)收入;

(6)市场;超市;

(7)天然;野生;

…

① http://www.xyxc.cn/article/show.asp?id=145

算法结束后输出的关键短语为

种植山野菜;收入;暖棚种植;收获;经济效益;

按照文献[20-21]介绍的基于统计信息抽取的关键词集合为

山野菜;野生;经济效益;种植;效益;

从上述实验结果中可以看出,由于基于统计信息的关键词抽取算法没有进行文章主题分析,某些高频词,例如"野生"并不一定能够确切反映文章的主题信息,同时算法抽取的描述同一主题的相似词过多,造成了信息冗余,例如"效益"和"经济效益"均描述同一主题.而基于词汇链抽取的关键短语能够更加全面地覆盖文中的主题信息,并且不同的关键词描述了不同的信息,不存在信息冗余问题.

表 1 列举了从 5 类测试语料集中任选两篇文档分别按词汇链和统计信息两种方式抽取的关键词的对比结果,将实验结果中每篇文档的关键词集合中反映相同主题的重复词语用"√"表示,将富含更多信息的短语用"◇"表示.例如从语料 Data Set 4 中的第 2 篇文档中抽取的短语"保鲜技术"就比"保鲜"富含更多信息,该短语表明文档是在描写一种使山野菜保持新鲜的技术而不像"保鲜"那么笼统,使读者不明白该文介绍的是关于"保鲜"的哪个方面,是"保鲜"的技术还是用途.

表 1 中,我们以序号代表从每个测试语料集中随机选择的文档号,以方式 A 代表基于统计信息抽取的关键词集合,以方式 B 代表基于词汇链抽取的关键词集合.

表 1　基于词汇链和基于统计信息抽取的关键词集合

序号	方式	文档关键词/短语集合				
Data Set 1						
1	A	警察√	刑侦√	侦查√	楷模	刑警√
	B	人民警察◇	学习楷模◇	信赖	刑侦能力◇	喝彩
2	A	形象	学习	荣誉	称号	保卫
	B	树立形象◇	学习精神◇	光荣称号◇	执法	保卫社会◇
Data Set 2						
1	A	救援√	海啸	抢险√	精神	调遣
	B	国际救援◇	海啸	调遣队伍◇	医生	无畏精神◇
2	A	赈灾√	救援√	香港	捐款	慈善机构◇
	B	赈灾活动◇	筹款	香港	捐赠	慈善机构◇
Data Set 3						
1	A	隐患	治理√	圆明园	意见	整治√
	B	暴露隐患◇	水治理◇	圆明园	公众意见◇	环保
2	A	污染	泄露	圆明园	紧缺	环保
	B	污染环境◇	泄露	圆明园	湖水紧缺◇	环保
Data Set 4						
1	A	山野菜	产品加工√	估算	投资	加工项目√
	B	食用山野菜◇	生态	绿色食品◇	投资	加工项目
2	A	山野菜	保鲜√	浸泡	欢迎	新鲜√
	B	山野菜	保鲜技术◇	清水浸泡◇	绿色	深受欢迎◇
Data Set 5						
1	A	足球机器人	竞赛√	比赛√	竞争√	技术创新◇
	B	足球机器人	技术创新◇	足球竞赛◇	竞争激烈◇	进步
2	A	足球机器人	决策√	决定√	发送	数据
	B	足球机器人	接收数据◇	决策系统◇	发送指令◇	通信

由表 1 可见,基于词汇链的方法能够抽取比单个词富含更多信息的短语,并且抽取的短语覆盖了不同的主题,解决了基于统计信息抽取时含有相同信息的冗余词过多,且没有全面反映文章主题的问题.

另外我们对上述语料的关键短语或关键词的抽取时间进行了对比,其对比的方法是基于统计信息的关键词抽取方法(表 2 中的方式 A)和基于词汇链的关键短语抽取方法(表 2 中的方式 B).分别记录测试语料中每个语料集包含的文档的关键词抽取时间,并将语料集内所有文档的抽取时间取平均值,将结果记录于表 2 中.

表 2　基于词汇链和基于统计信息的关键词抽取时间

方式	抽取时间/s				
A	3.13	5.82	4.57	4.27	6.41
B	4.51	7.24	5.65	4.91	8.36

由表 2 可见,基于词汇链的关键短语抽取方法的抽取时间略高于基于统计信息的关键词抽取方法

的抽取时间.其原因正如我们在文中第5节的分析,基于词汇链的关键短语抽取方法的时间复杂度为$O(n^2)$,其与基于统计信息的关键词抽取方法的时间复杂度相当,然而由于基于词汇链的关键词短语抽取方法的步骤多于基于统计信息的关键词抽取方法,例如候选中心词聚类步骤,因此其抽取时间略高于基于统计信息的关键词抽取方法.

将上述5类语料进行混合后以基于词汇链和基于统计信息的方法分别抽取不同数目的关键短语并计算准确率,其准确率为
(算法返回的短语集合与人工标注的短语集合中相同短语的数目)/(算法返回的短语集合中的短语数),结果如图2所示.

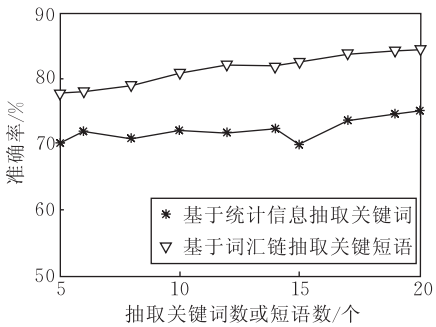

图2 关键词抽取准确率对比图

由图2可见,采用统计信息进行关键词抽取时准确率曲线上下跳动的幅度很大,这说明基于统计信息的关键词抽取方法不能观测到文章的主题分布,抽取的关键词不能完全覆盖文章的多个主题信息.而本文介绍的算法(基于词汇链的关键短语抽取算法)能够很好地避免上述问题,其根据文章的主题信息,选取最能够代表文章主题的关键短语进行抽取,很好地覆盖了文章的多个主题线索,即使抽取少量短语时对文章主题的覆盖性能也很强,同时对属于相同主题线索的多个关键短语只选取其中最具描述能力的短语输出,避免了信息冗余,因此准确率曲线基本呈上升趋势.

我们以基于词汇链和基于统计信息两种方法抽取的关键词作为聚类特征用于文本聚类.聚类是将文档集合按文档间的内在相似度进行划分,使同类文档的相似度较高而不同类文档的相似度较小.如果本算法(基于词汇链的关键短语抽取算法)能够获得更加准确的聚类结果,则说明此算法抽取的特征或关键词能够比统计特征在更深的层次上反映文档间的关系.分别采用基于词汇链和基于统计信息两种方法抽取5、10、15个词作为文档的特征词,然后采用层次聚合聚类对测试语料集进行聚类[31],本文采用*Rand*值来计算聚类的准确率[32].

在*Rand*值评价体系中,测试数据集中的每两个数据被视为一个点对.假设测试数据集共包含p篇文档,则测试数据集含有$p(p-1)/2$个点对,同时以4种情况来标识聚类结果:

(a)点对在测试语料中位于同一类别中,在聚类结果中位于同一类别中.

(b)点对在测试语料中位于同一类别中,在聚类结果中位于不同类别中.

(c)点对在测试语料中位于不同类别中,在聚类结果中位于不同类别中.

(d)点对在测试语料中位于不同类别中,在聚类结果中位于同一类别中.

设满足上述4种情况的点对数目分别为a、b、c、d.可见上述a和c分别代表了正确划分的点对数目,则*Rand*值即如式(11)所示.

$$Rand=\frac{a+c}{p\times(p-1)/2} \tag{11}$$

本文以2004年的863语料作为基准的聚类数据集[9],其采用中图分类法体系,但不包括难以辨别的"T 工业技术"和"Z 综合性图书"两类.此语料包含了36个类别,每个类别包含100篇文档.对上述聚类语料,从每个类别中随机选择50篇文档,这样即可获得一个包含36个类别,每个类别中含有50篇文档的聚类测试语料,重复上述方法4次即可构成4个标准的聚类测试语料集,其聚类结果如表3所示.

表3 基于词汇链和基于统计信息得到的特征的聚类结果(其中5,10,15为抽取的特征数)

方式	聚类结果					
	5	10	15	5	10	15
	Data Set 1			Data Set 2		
A	79.49	82.46	80.77	71.23	71.89	75.62
B	81.42	82.57	83.69	72.03	74.71	76.55
	Data Set 3			Data Set 4		
A	82.41	82.13	84.52	76.91	78.03	79.37
B	82.41	83.19	84.93	76.91	79.45	80.67

表3中我们以方式A代表以基于统计信息抽取的关键词集合作为特征向量得到的聚类结果;以方式B代表以基于词汇链抽取的关键词集合作为特征向量得到的聚类结果.

由表3可见,以基于词汇链抽取的关键词作为

聚类特征进行聚类后,聚类结果更加准确.这说明基于词汇链的关键词抽取方法能够深入到文本的语义一级,抽取的关键词能够更好地反映文本的主题信息,挖掘出文本的隐含信息.同时可以发现,在基于统计信息的抽取方法获得的聚类结果中,随着抽取词数的增多,聚类准确率却不一定上升,而在基于词汇链的抽取方法获得的聚类结果中,聚类准确率随抽取词数的增多一直上升.这是因为基于统计信息的抽取方法中抽取的关键词可能存在一些与文档描述的主题信息无关的高频词,当这些词作为聚类特征时显然会降低聚类结果的准确性,而基于词汇链的关键词抽取方法却通过构造词汇链对文档的主题进行了充分的分析,因此抽取的关键词能够很好地覆盖文档的主题信息,其聚类准确率随抽取词数的增多在逐渐上升.

7　结　论

关键词是对文章主题信息的精炼,好的关键词能够有效提高读者的阅读速度,加深读者对文章的理解.本文提出了一种基于词汇链的关键短语抽取算法,此算法能够根据文章的主题分布动态确定输出短语的数目,使短语能够全面覆盖文章描述的多个主题线索,并且不同短语描述不同的主题信息,信息冗余度小.实验表明这种基于词汇链的方法抽取的关键短语覆盖的信息量要远优于基于统计信息抽取的关键词,其能够深入到文章的语义一级,挖掘出文章所包含的深层主题信息,达到了一个比较理想的效果.

参　考　文　献

[1] Bracewell D B, Ren F, Kuriowa S. Machine learning techniques for business blog search and mining. Expert Systems with Applications, 2008, 35(3): 581-590

[2] Zhu Huafei, Bao Feng. Continuous keyword search on multiple text streams//Proceedings of the IEEE International Conference on Communications. Glasgow, England, 2007: 1336-1341

[3] Vagelis H, Oscar V, Michail V, Philip S Y. Continuous keyword search on multiple text streams//Proceedings of the 15th ACM International Conference on Information and Knowledge Management. Arlington, Virginia, USA, 2006: 802-803

[4] Yang Weidong, Shi Baile. Schema-aware keyword search over XML streams//Proceedings of the 7th IEEE International Conference on Computer and Information Technology. Fukushima, Japan, 2007: 29-34

[5] Xu Yan, Li Jin-Tao, Wang Bin, Sun Chun-Ming, Zhang Shen. A study on feature selection algorithm in text categorization technology. Journal of Computer Research and Development, 2008, 45(4): 596-602(in Chinese)
(徐燕,李锦涛,王斌,孙春明,张森. 文本分类中特征选择的约束研究. 计算机研究与发展, 2008, 45(4): 596-602)

[6] Chang Hischeng, Hsu Chiunchieh. Using topic keyword clusters for automatic document clustering//Proceedings of the 3rd International Conference on Information Technology and Applications. Sydney, Australia, 2005: 419-424

[7] Akiko A. An information-theoretic perspective of tf-idf measures. Information Processing and Management, 2004, 39(1): 45-65

[8] Ye C, Divakaran L. A maximum entropy approach to feature selection in knowledge-based authentication. Decision Support Systems, 2008, 46(1): 388-398

[9] Liu Tao, Liu Bing-Quan, Xu Zhi-Ming, Wang Xiao-Long. Automatic domain-specific term extraction and its application in text classification. Acta Electronica Sinica, 2007, 35(2): 328-332(in Chinese)
(刘桃,刘秉权,徐志明,王晓龙. 领域术语自动抽取及其在文本分类中的应用. 电子学报, 2007, 35(2): 328-332)

[10] Witten I H, Paynter G W, Frank E, Gutwin C, Nevill-Manning C G. KEA: Practical automatic key-phrase extraction//Proceedings of the 4th ACM Conference on Digital Libraries. Berkeley, CA, USA, 1999: 254-255

[11] Turney P D. Learning to extract key phrases from text. National Research Council, Canada: NRC Technical Report ERB-1057, 1999

[12] Li Sujian, Wang Houfeng, Yu Shiwen, Xin Chengsheng. Research on maximum entropy model for keyword indexing. Chinese Journal of Computers, 2004, 27(9): 1192-1197(in Chinese)
(李素建,王厚峰,俞士汶,辛乘胜. 关键词自动标引的最大熵模型应用研究. 计算机学报, 2004, 27(9): 1192-1197)

[13] Yang Wenfeng. Chinese keyword extraction based on max-duplicated strings of the documents//Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Tampere, Finland, 2002: 439-440

[14] Wu Xiaoyuan, Alvaro Bolivar. Keyword extraction for contextual advertisement//Proceedings of the 17th International Conference on World Wide Web. Beijing, China, 2008: 1195-1196

[15] Liu Yuan-Chao, Wang Xiao-Long, Liu Bing-Quan, Zhong Bin-Bin. The clustering analysis technology for information retrieval. Journal of Electronics and Information Technology, 2006, 28(4): 606-609(in Chinese)
(刘远超,王晓龙,刘秉权,钟彬彬. 信息检索中的聚类分析技术. 电子与信息学报, 2006, 28(4): 606-609)

[16] Liu Qun, Li Su-Jian. Word similarity computing based on Hownet. Computational Linguistics and Chinese Language Processing, 2002, 17(2): 59-76(in Chinese)
(刘群, 李素建. 基于《知网》的词汇语义相似度计算. 中文计算语言学期刊, 2002, 17(2): 59-76)

[17] Victoria J H, Jim A. Hierarchical word clustering-automatic thesaurus generation. Neurocomputing, 2002, 48(1-4): 819-846

[18] Sven M, Jorg L, Hermann N. Algorithms for bigram and trigram word clustering. Speech Communication, 1998, 24(1): 19-37

[19] Bekkerman R, El-Yaniv R, Tishby N, Winter Y. Distributional word clusters vs words for text categorization. Journal of Machine Learning Research, 2003, 3(7-8): 1183-1208

[20] Anette H. Combining machine learning and natural language processing for automatic keyword extraction [Ph. D. dissertation]. Department of Computer and Systems Sciences, Stockholm University, Stockholm, Sweden, 2004

[21] Anette H, Beata B M. A study on automatically extracted keywords in text categorization//Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. Sydney, Australia, 2006: 537-544

[22] Gonenc E, Ilyas C. Using lexical chains for keyword extraction. Information Processing and Management, 2007, 43(6): 1705-1714

[23] Gan Kok Wee, Wong Ping Wai. Annotating information structures in Chinese texts using Hownet//Proceedings of the 2nd Workshop on Chinese Language Processing: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics. HK, 2000: 85-92

[24] Lu Zhi-Mao, Liu Ting, Li Sheng. The research progress of statistical word sense disambiguation. Acta Electronica Sinica, 2006, 34(2): 333-343(in Chinese)
(卢志茂, 刘挺, 李生. 统计词义消歧的研究进展. 电子学报, 2006, 34(2): 333-343)

[25] Chen Qing-Cai, Wang Xiao-Long. A word vector based quantization model of Chinese word sense. Journal of Computer Research and Development, 2001, 38(2): 207-212(in Chinese)
(陈清才, 王晓龙. 一种基于词矢量的汉语语义量化模型. 计算机研究与发展, 2001, 38(2): 207-212)

[26] Levitin Anany V. Introduction to the Design and Analysis of Algorithms. 2nd Edition. Massachusetts, USA: Addison Wesley, 2002

[27] Li Su-Jian. Research of relevancy between sentences based on semantic computation. Computer Engineering and Applications, 2002, 38(7): 75-76(in Chinese)
(李素建. 基于语义计算的语句相关度研究. 计算机工程与应用, 2002, 38(7): 75-76)

[28] Xu Yong-Dong, Xu Zhi-Ming, Wang Xiao-Long, Liu Yua-Chao, Liu Tao. Using multiple features and statistical model to calculate text units similarity//Proceedings of the 2005 International Conference on Machine Learning and Cybernetics. Guangzhou, China, 2005: 18-21

[29] Liu Yuan-Chao, Wang Xiao-Long, Xu Zhi-Ming, Liu Bing-Quan. Mining construction rules of Chinese key phrase based on rough set theory. Acta Electronica Sinica, 2007, 35(2): 371-374(in Chinese)
(刘远超, 王晓龙, 徐志明, 刘秉权. 基于粗集理论的中文关键词短语构成规则挖掘. 电子学报, 2007, 35(2): 371-374)

[30] Sun Mao-Song, Zuo Zheng-Ping, Tsou B K. Part-of-speech identification for unknown Chinese words based on k-nearest neighbors strategy. Chinese Journal of Computers, 2000, 23(2): 166-170(in Chinese)
(孙茂松, 左正平, 邹嘉彦. 基于 k-近似的汉语词类自动判定. 计算机学报, 2000, 23(2): 166-170)

[31] Qu Jun, Jiang Qing-Shan, Weng Fang-Fei, Hong Zhi-Ling. A new hierarchical clustering based on overlap similarity measure//Proceedings of the 8th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, Qingdao, China, 2007: 905-910

[32] Wang Ling, Bo Lie-Feng, Jiao Li-Cheng. Density-sensitive semi-supervised spectral clustering. Journal of Software, 2007, 18(10): 2412-2422(in Chinese)
(王玲, 薄列峰, 焦李成. 密度敏感的半监督谱聚类. 软件学报, 2007, 18(10): 2412-2422)

LIU Ming, born in 1981, Ph. D. candidate. His research interests include clustering analysis and text digging.

WANG Xiao-Long, born in 1955, professor, Ph. D. supervisor. His research interests include natural language processing and text digging.

LIU Yuan-Chao, born in 1971, associate professor. His research interests include clustering analysis and artificial intelligence.

Background

With advance of network technique, people need to contact much more information every day. How to acquire useful information from a mass of data is a valuable research work. Keyword is a simple information description instrument. In order to know topics of documents, users only need to read smaller keywords or key-phrases. However, recent keyword extraction methods are often based on statistics. Many words extracted by these methods can't reflect topics of documents. It also exists problems of information redundancy among keywords and difficult to confirm extraction number of keywords.

The research in this paper is supported by three foundations. They are National Natural Science Foundation of China (60435020) and National High Technology Research and Development Program (863 Program) of China (2006AA01Z197, 2007AA01Z172).

The first project and the second project are both to research how to improve traditional retrieval systems. The first project uses NLP techniques to improve retrieval efficiency. The second project develops retrieval system for special domain such as finance. In these projects, keyword extraction method is used to construct index to improve retrieval precision and recall.

The third project develops techniques for large-scale documents clustering. The features which reflect topic of documents obviously can measure similarities among documents well. So keyword extraction method is used to select features to construct feature vector to improve clustering precision. Besides, when scale of documents augments, dimension number of feature vector also increases rapidly. The huge feature vector increases clustering time greatly. Keyword extraction method is used to remove features which are irrelative to topic of documents. In this project, keyword extraction is also used to form cluster label to make users easily know what the cluster reflects.

Keyword extraction based on lexical chain, which is proposed in this paper, first constructs lexical chains to reflect topic of documents. After that, it selects keywords or keyphrases by lexical chains. This extraction method not only can accelerate reading speed of users but also can apply in many fields as pretreatment approach, such as information retrieval and document clustering.