结合语义扩展度和词汇链的关键词提取算法

刘端阳 王良芳
(浙江工业大学计算机科学与技术学院 杭州 310023)

摘 要 针对影响关键词提取质量的一词多义现象、同义词现象以及文章主题准确全面表达的难点,提出了一种基于语义的关键词提取算法KESELC,利用《同义词词林》语义词典和统计信息计算语义相似度和相关度,进而得出语义扩展度及其计算方法,将语义扩展度和词汇链方法相结合,对文本分别作预处理、多义词词义消歧、同义词合并、词汇链构建、有效特征选取及对权重综合计算的处理,提取出的关键词不仅避免了同义词冗余表达,而且较准确全面地覆盖文本的主题。通过实验对比分析,验证了基于KESELC的方法比基于TFIDF的方法以及基于词汇链的方法具有较优的提取效果,具有一定的实际应用价值。

关键词 同义词词林,语义扩展度,词汇链,关键词提取,语义分析
中图法分类号 TP391 **文献标识码** A

Extraction Algorithm Based on Semantic Expansion Integrated with Lexical Chain

LIU Duan-yang WANG Liang-fang
(College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract For the difficulties that affect the quality of keywords extraction, such as the phenomenon of polysemy, synonyms as well as the accurate and comprehensive expression of the subjects in the text, a method named KESELC based on the semantics of keyword extraction was proposed. By calculating semantic similarity and semantic relevancy based on the tongyici cilin and statistical information, then the concept of semantic expansion and its calculation method were proposed. By combining semantic expansion with lexical chain, it made the text processing in terms of preprocess, polysemy disambiguation, synonym mergence, the construction of lexical chains, feature selection and improvement of weights computation. The extracted keywords not only avoid a redundant expression, but also cover the subjects of the article accurately and comprehensively. The experimental results show that the method of keyword extraction based on KESELC has better performance than the ones based on TFIDF and Lexical chain, and has a certain practical value.

Keywords Tongyici cilin, Semantic expansion, Lexical chain, Keyword extraction, Semantic analysis

1 引言

随着信息时代的发展,大量的文本信息以数据化的形式存在,许多领域都不断产生海量数据,因此如何在海量的信息中,快速并准确地提取对读者有用的信息变得越来越重要。关键词标注方法就是一种解决上述问题的有效手段,关键词是对文章主题信息的精炼,有助于提高读者的阅读速度,使读者快速掌握该文本是否是自己所感兴趣的信息,加深读者对文章的理解。关键词提取一直是文本挖掘领域的主要研究问题,同时关键词提取技术也是自动文摘[1]、文本分类[2]、话题检测[3]等其他自然语言处理研究的基础工作,并起着十分重要的作用。然而目前除了学术刊物论文一般已赋予关键词外,绝大多数的其他文本信息还尚未提供关键词,同时手工抽取关键词既耗时又具有较强的主观性,因此关键词自动提取是文本数据挖掘领域的一个研究热点。

国内外相关文献相继提出了很多关键词自动提取方法,主要可以概括为3类:基于统计信息的关键词自动提取、基于机器学习的关键词自动提取和基于语义分析的关键词自动提取。其中,基于统计信息的关键词提取方法,如基于词频[4]等方法,虽然简单易行、适用性较强,但关键词提取的准确率往往比较低,会出现一些高频而重要性低的词语被误选上等情况。基于机器学习的关键词提取方法,如基于Naïve Bayes的方法[5]、基于SVM的方法[6]等也被较多地采用。该类方法的基本思想将关键词提取问题视为二元分类问题,在训练阶段根据提取关键词的特征构建关键词分类模型,然后在分类阶段基于模型从文档中抽取词并判断该词是否为关键词。但该类方法一方面面临训练样本标注的瓶颈问题[7],即已标引的样本相对整个样本空间上的数据分布是有限的,并且很难达到整个样本空间的数据,少量的已标引样本和大量的未标引样本的不平衡,难以训练出好的分类器;另一方面提取关键词虽然不限制于语种,但训练出的分类器只有在对与训练语料领域类似的文章进行关键词提取时才能取得较不错的效

到稿日期:2013-01-17 返修日期:2013-06-14 本文受国家自然科学基金(61202204)资助。
刘端阳(1975—),男,博士,副教授,主要研究方向为数据挖掘、分布式计算等,E-mail:ldy@zjut.edu.cn;王良芳(1988—)男,硕士,主要研究方向为数据挖掘、人工智能。

果，并且分类器训练过程中还可能存在过拟合的问题。

近几年来，基于语义的关键词自动提取方法从语义的角度分析词语的关键性，与人们的感知逻辑思维比较符合，得到了国内外研究者们广泛的关注。这类方法主要运用语义词典（如英文的 WordNet，中文的《同义词词林》、《知网》等）获取词汇间的语义知识来提取文本关键词，文献[8]采用基于 WordNet 计算词语相关度来考虑词语间的内聚性，将内聚性、首次出现位置以及词频作为特征加权的主要考虑因素。文献[9]提出了一种基于词相似度词典的方法，即采用 TFIDF、节点度等特征，使用 Weka 训练建立朴素贝叶斯预测模型获取关键词。文献[10]提出一种基于 HowNet 汉语间语义关系来计算语义相关度的方法。文献[11]提出了一种在语义网环境下结合语义相似度和相关度的语义扩展度概念。文献[12]提出将语义相关的词语组成词汇链，进行语义特征分析来获取表达文章主题的关键词。但已有的基于语义的方法还无法很好地解决多义词词义消歧、同义词冗余表达、构建的词汇链不能准确表达文本语义结构等问题，影响了关键词提取的质量。同时，在已有的基于语义的方法中还没有将由语义相似度与语义相关度构建出的语义扩展度和词汇链相结合的方法。相对来讲，国内的中文语义词典不如国外的语义词典应用广泛而且数量相对较少，同时容易获取用于研究使用的也只有《同义词词林》，并且还没有一个系统的基于《同义词词林》语义分析的关键词提取方法。

因此，本文提出一种基于《同义词词林》语义词典的语义扩展度和词汇链方法相结合的关键词提取算法 KESELC（Keyword Extraction Based on Semantic Expansion Integrated With Lexical Chain）。首先对候选关键词进行词义消歧，根据《同义词词林》中词汇间的语义关系和统计信息，计算出词汇间的语义相似度和语义相关度，进而计算语义扩展度，构建词汇链，综合考虑词汇所在词汇链的强度、词频、区域位置等特征信息，最终按权值高低输出关键词。实验结果表明，与已有算法相比，该算法能较好地提高关键词提取的质量。

2 相关知识介绍

2.1 同义词词林简介

由梅家驹等[13]编纂的《同义词词林》是一部汉语分类词典，其按树状层次结构把词条组织起来。由于《同义词词林》著作时间较久远，导致一方面某些词汇成为生僻词，另一方面新词也没有加入到词典中，限制了其语义处理的功能。因此哈尔滨工业大学信息检索实验室利用词语相关资源，完成了一部具有汉语大词表的“哈工大信息检索研究室同义词词林扩展版”[14]。

《同义词词林》语义词典具有 5 层结构，词典中每个词对应相应的编码，每个编码由 5 层代码和一位标记组成，即编码 $Code_i$ 表示为 $Code_i = A_{i1}A_{i2}A_{i3}A_{i4}A_{i5}F_i$。例如：“Gb12A01＝ 尊敬 崇敬 敬重 敬爱 尊崇”，其中“Gb12A01＝”是编码，“尊敬”、“崇敬”等都是该编码所对应的词语。编码位按从左到右的顺序排列，编码层次分支越靠右，表示词语间语义越相似。第 8 位标记具有“＝”、“＃”、“@”3 种不同表示形式。其中“＝”表示词语同义；“＃”表示词语相关；“@”表示词语独立，其在词典中既没有同义词，也没有相关词。

《同义词词林》编码简单，具有丰富的语义知识，并且还没有一个基于同义词词林的语义扩展度和词汇链方法相结合的文本关键词提取方法。因此通过分析《同义词词林》的特性，结合语义扩展度和词汇链方法，提出了一个完整的基于语义的文本关键词提取新方法。

2.2 语义相似度和语义相关度简介

语义相似度指出现在文章不同位置的两个词语可以互相替换而不改变文章句法语义结构的程度，如“领导-上司”。语义相关度指出现在文章不同位置的两个词语间的相关联程度，这两个词语可能不存在相似关系，但可能通过某种关联具有相关关系，如“领导-员工”，相似度较小，但相关度却很大。一般来讲，如果两个词语间的相似度高，那么相关度也高；然而如果两个词语间的相关度高，两者之间的相似度却不一定高[11]。上述例子中表明，语义相似度和语义相关度具有一定的蕴涵关联，相似性不等同于相关性，相关性和相似性也不是互斥关系。

因此，提出了结合语义相似度和相关度的语义扩展度的概念，将具有一定程度的语义相似和相关的词语组成在一起共同来表达文章的某一主题内容。

2.3 词汇链简介

词汇链是指在文本中一系列相似词或相近词共同组成的链式集合体，词汇链共同表达了文章描述的主题信息，其中每条链表达某一主题信息，选取强度较高的词汇链，并从这些链中提取能充分表达该链所述信息的词语作为候选关键词语。

词汇链构建的主要思想就是对文本预处理和词义消歧处理后的候选词汇集逐项进行选择，计算词语与每个词汇链的词义相似度，然后选择与当前处理的词具有最大相似度的词汇链作为该词所属的链，并将该词插入到此词汇链中，最后根据词汇链的特征信息计算权重，选择强度较高的词汇链来表达文档所描述的主题信息。因此词汇链对于提取关键词有着重要意义。

3 结合语义扩展度与词汇链的关键词提取算法

中文文本中多义词现象存在较为普遍，但在文中具体位置的词汇在该上下文语境中只含有一种词义，为正确表达文章语义结构，必须实现多义词词义消歧，这也是提高关键词提取质量的基础工作。同时，提取出的关键词应该尽量全面确切地表达文章的主题内容而又要避免同义词的冗余出现，即避免表达同一主题的相似或相关的词语重复出现。

为了解决以上问题，提出一种基于《同义词词林》语义词典的语义扩展度和词汇链方法相结合的关键词提取算法 KESELC。首先对文本进行预处理，采用应用较广并且分词效果不错的中国科学院计算技术研究所汉语词法分析系统 ICTCLAS[15]对文档进行分词和词性标注；然后根据词性进行过滤，只选取名词、动词作为候选词。同时利用停用词表对一些明显不能成为关键词的停用词进行过滤；统计过滤后的候选词汇集中词汇的频率和位置信息。然后对预处理后的候选关键词进行词义消歧，根据《同义词词林》中词汇间的语义关系和统计信息，计算出词汇间的语义相似度和语义相关度，综合考虑两者计算语义扩展度，构建词汇链，对表达同一概念的同义词进行合并，综合考虑词汇所在词汇链的强度、词频、区域位置等特征信息，最终按权值高低输出关键词。KESELC 算

法流程如图1所示。

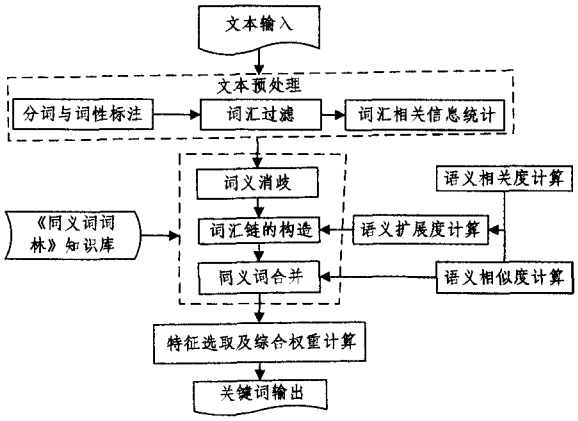

图1 KESELC算法流程

KESELC关键词提取算法主要贡献有以下3个方面：

(1)针对目前对《同义词词林》研究除了语义相似度计算外，其他基于《同义词词林》的语义研究还较少，综合考虑《同义词词林》语义知识和统计信息，提出了基于《同义词词林》的语义相关度计算方法。

(2)根据基于《同义词词林》语义词典的语义相似度和语义相关度，提出了语义扩展度及其计算方法，利用基于《同义词词林》的语义扩展度构建词汇链。

(3)在《同义词词林》语义词典的基础上，利用《同义词词林》语义结构特点，根据词汇在文章上下文和语义词典中出现的情况，计算词汇间的共现率，确定多义词的具体词义，从而为提高提取质量奠定了基础。依据词汇间的语义相似度，将一定程度上比较相似的同义词进行合并处理，有效避免了关键词输出结果中冗余表达等情况的发生。

(4)在特征选取及综合权重计算阶段，针对语义扩展度和词汇链相结合的优势，提出既考虑词频、区域位置特征又结合考虑词汇链强度特征的综合权重计算。

3.1 语义相似度的计算

研究者们提出了多种词汇相似度的计算方法，如利用语义词典(同义词词林、知网、WordNet等)进行计算。本文采用田久乐[16]提出的基于词典编码距离的词汇语义相似度计算方法。主要思想为基于同义词词林的结构，判断需进行相似度计算的两个词汇所对应的编码在哪一层分支，根据两个义项的语义距离，计算出相似度，距离越近，相似度越高。语义相似度计算公式如下：

$$\mathrm{Sim}(w_1,w_2)=d\cdot\left(\frac{n-k+1}{n}\right)\cdot\cos\left(n\cdot\frac{\pi}{180}\right) \tag{1}$$

式中，Sim为语义相似度($0<\mathrm{Sim}<1$)；d为系数，由两个词汇所对应的编码在哪一层分支所决定；n为分支层节点的总个数；k为分支间的距离。

3.2 语义相关度的计算

语义相关度指出现在文章不同位置的两个词语间的相关联程度。相关度的度量通常带有一定的主观色彩，一种方法是利用与用户交互的相关反馈，另一种方法建立在语料库或统计词典的基础上，相对比较客观[11]。

本文提出了利用《同义词词林》词汇间语义关系和统计的方法来计算相关度。一般来说，如果两个词在同一文章中同时出现的次数越多，则两个词之间的相关度越大。相关度计算方法的主要思想如下：首先查询要进行相关度计算的词w_1和词w_2在《同义词词林》语义词典中分别对应的编码$Code_1$和$Code_2$，如果$Code_1$等于$Code_2$，并且其中第8位标记位为“#”，则两个词之间的相关度为1；如果$Code_1$等于$Code_2$，并且其中第8位标记位为“=”，则两个词之间的相关度为0.85；否则，分别统计这两个词文章中同时出现的次数和各自单独出现的次数，然后将统计信息代入式(2)计算得到两个词之间的相关度。

$$rel(w_1,w_2)=\frac{count(w_1,w_2)}{\min(count(w_1),count(w_2))} \tag{2}$$

式中，$count(w_1,w_2)$为两个词同时出现的次数，$\min(count(w_1),count(w_2))$为词$w_1$和词$w_2$单独出现次数的较小值。式(2)中选用单独出现次数的较小值作为分母计算，主要考虑实际情况中可能会出现一个词为高频词、另一个词为低频词的情形，在这种情形下，无论选用两个词单独出现次数的较大值或单独出现次数的平均值参与计算，均会造成较大的误差，这时应该采用出现次数较小的低频词参与计算才能获得较准确的结果。

同时，通过实验分析发现，与给定的专家值相比，计算结果普遍偏小，但数据集的计算结果整体趋势是一致的，只要对公式进行优化处理，一定程度上提高整体的幅度就可得到比较理想的效果。通过实验发现，如果对结果开平方根，那么计算结果和专家值就比较接近了，则公式(2)可以改进为：

$$rel(w_1,w_2)=\sqrt{\frac{count(w_1,w_2)}{\min(count(w_1),count(w_2))}} \tag{3}$$

综上所述，两个词之间的相关度计算公式为：

$$rel(w_1,w_2)=\begin{cases}1, \text{if } Code_1=Code_2 \text{ and } F_1=F_2='\#'\\ 0.85, \text{if } Code_1=Code_2 \text{ and } F_1=F_2='='\\ \sqrt{\dfrac{count(w_1,w_2)}{\min(count(w_1),count(w_2))}}, \text{else}\end{cases} \tag{4}$$

3.3 语义扩展度的计算

基于《同义词词林》语义词典，根据词典中词语的语义关系和蕴涵关联得到语义相似度计算式(1)和语义相关度计算式(4)，结合相似度和相关度，提出语义扩展度的计算公式如下：

$$\exp(w_1,w_2)=\alpha\times\mathrm{Sim}(w_1,w_2)+(1-\alpha)\times rel(w_1,w_2) \tag{5}$$

式中，α为在语义扩展中调节相似度和相关度所占比重的参数，本实验中α取为0.5。

3.4 词义消歧

同一个词在不同的上下文中可以表达不同的语义，即词形相同而语义不同的多义词现象。例如：“一尘不染”一词，在《同义词词林》语义词典中出现多行，分别出现在“Ee41A01=廉洁 清廉 廉正 廉洁奉公 两袖清风 一尘不染 一身清白”和“Ef12A01=清洁 干净 卫生 整洁 清爽 一尘不染 窗明几净”中，很明显通过其他词可以看出，前一个“一尘不染”用来形容人，表示为官清廉的意思，而后一个“一尘不染”用来形容环境场所，表示为干净的意思。判别多义词的具体词义在很大程度上是由该词所在的上下文语境所决定的。文献[17]在运用WordNet语义词典解决英文多义词分歧的问题上，认为如果某个多义词的其中一条注释中的词在该词的上下文中出现率较高，即可认为这条注释是该多义词在当前语境中所属的词义。因此针对中文文本多义词分歧现象较为普遍这一问题，

提出了基于《同义词词林》来计算词汇间的共现率，从而确定该多义词的具体词义。

在同义词词林中，属于同一编码甚至属于第5层中的词语都是词义相近的。如果其中某个词是多义词，那么其他词就可以视为该多义词的注释，采用式(6)计算词汇间的共现率。

$$T(w_1,w_2)=\log_2\left\{\frac{\sum_s P(w_1,w_2)}{P(w_1)\cdot P(w_2)}\right\} \tag{6}$$

式中，w_1 表示多义词；w_2 表示注释中的词；s 表示语义范围，即以“。”、“！”、“？”等作为分隔符的区域。文献[6]中认为某个词的影响范围近似为该多义词的前两词和后两词，但5个词范围远远不能满足中文文本的要求，所以提出了更贴切中文文本特点的语义范围概念。其中 P 表示在文本中的出现概率，概率计算公式为：

$$P(w)=\frac{f(w)}{F} \tag{7}$$

式中，$f(w)$ 表示词汇 w 的词频；F 表示所有词汇总词频。所以式(6)可转换如下：

$$T(w_1,w_2)=\log_2\left\{\frac{\sum_S F*f(w_1,w_2)}{f(w_1)*f(w_2)}\right\} \tag{8}$$

由式(8)可得出，w_1，w_2 同时出现得越频繁，$P(w_1,w_2)$ 越大，$T(w_1,w_2)$ 也就越大，表明 w_1，w_2 的重复性越高，越能代表多义词 w_1 的含义。反之亦然。通过以上处理，就可以确定该多义词在当前上下文语境中的准确含义，从而解决多义词歧义问题。

3.5 词汇链的构造

Morris 和 Hirst[18]最早提出了构建词汇链的方法，该算法针对英文文本，利用 WordNet 知识库并仅选择出现在 WordNet 中的所有名词作为候选词，提出了词汇链计算模型以及词汇链的生成算法。

与之不同，提出了基于《同义词词林》语义词典来确定中文词汇间关系并构建词汇链的方法。选取的候选词为《同义词词林》中收录的名词、动词，从中逐项选择候选词，计算词与词汇链的语义扩展度后，将该词加入到对应的词汇链中。具体算法如下：

Step1 预处理文档集，包括分词、词性标注和词过滤，并对每个词的特征频率 TF 和文档频率 DF 进行统计。

Step2 由于有些领域词汇未被《同义词词林》收录但出现频率较高，这些词汇成为关键词的概率较高，因此将特征频率 TF 大于指定阈值 ε（一般 ε 取值为3）的未登录词单独归为词汇链 L_0。

Step3 选择文本中预处理后的词 $w_1,w_2,\cdots,w_n$ 作为候选词汇集，并取 w_1 构建初始词汇链 L_1。

Step4 对候选词汇集的词 $w_i(i\in[2,n])$ 依次进行提取，依次计算它与除词汇链 L_0 之外的词汇链 $L_j(j\in[1,m])$ 的语义扩展度 $\exp(w_i,L_j)$，即该词与某词汇链 L_j 中所有词的词义扩展度最大值作为词与该词汇链的扩展度；然后对每个 $S(w_i,L_j)$ 进行比较，选取其中的扩展度最大值作为该词与所有词汇链的扩展度 $\exp(w_i,L)$，即

$$\begin{aligned}\exp(w_i,L)&=\max_{j=1,2,\cdots,m}\exp(w_i,L_j)\\&=\max_{j=1,2,\cdots,m}\left[\max_{k=1,2,\cdots,n_j}\exp(w_i,w_{jk})\right]\end{aligned} \tag{9}$$

式中，n_j 为词汇链 L_j 中包含词汇的个数；m 为词汇链的条数；w_{jk} 为词汇链 L_j 中第 k 个词汇。词汇间语义扩展度 $\exp(w_i,w_{jk})$ 按式(5)计算，$\exp(w_i,L_j)$ 表示词汇 w_i 与词汇链 L_j 的语义扩展度。

Step5 将语义扩展度 $\exp(w_i,L_j)$ 的最大值和预设的阈值 σ 作比较，如果 $\exp(w_i,L_j)$ 最大值大于阈值 σ，就把词 w_i 加入到对应的词汇链 L_j 中。

Step6 如果语义扩展度 $\exp(w_i,L_j)$ 最大值小于阈值 σ，就创建一个新词汇链，并把词 w_i 加入到该新建的词汇链中。

Step7 对全部候选词汇依次进行计算，重复 Step3－Step6，直到全部词汇计算完毕。

在上述算法中，通过观察发现，扩展度阈值 σ 选择得越大，构建的词汇链数目就越多。反之，阈值 σ 选择得越小，构建的词汇链数目就越少。

3.6 同义词合并

针对较为普遍存在的同义词现象，即文章作者很有可能选择不同的词汇来表达同一概念，如“高兴”和“开心”，“中国”和“中华人民共和国”等，不仅由于重复计算而产生不必要开销，而且会造成关键词在输出结果中冗余表达等问题，严重影响关键词提取质量；同时《同义词词林》语义知识库比其他知识库拥有更丰富且易于语义理解的同义词组的优势，因此本文提出了基于《同义词词林》的同义词合并处理，在《同义词词林》的基础上，根据词汇间的相似度，将大于预设的相似度阈值 t 的同义词进行合并，并且对合并的词频作加权处理，一般 t 取为0.9。同义词合并处理流程如图2所示。

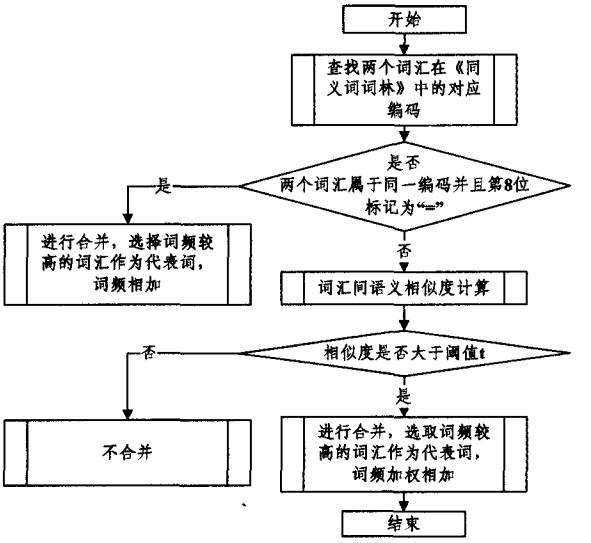

图2 同义词合并处理流程

《同义词词林》中，若两个词汇属于同一编码并且第8位标记为“=”的词汇表示两个词汇词义相同，相似度为1，当然可以进行合并。若两个词汇不出现在同一行，则根据式(1)进行词汇间的相似度计算，判断是否为同义词，即相似度是否大于阈值 t，判定是否可以进行合并。若可以合并，则选择词频高的词作为代表词，最终的词频为加权后的词频，计算公式如下：

$$f=f_1+\mathrm{Sim}(w_1,w_2)\cdot f_2 \tag{10}$$

式中，f_1 为出现词频较高的词汇词频，f_2 为出现词频较低的词汇词频。

如“飞行员”的词频是 f_1，“宇航员”的词频是 f_2（其中 $f_1>f_2$），那么“飞行员”最终的词频是 $f_1+\mathrm{Sim}(w_1,w_2)\cdot f_2$。

其中 Sim(w_1,w_2)为“飞行员”和“宇航员”的语义相似度。

3.7 特征选取及权重计算

特征信息决定候选词的权值。因此,特征选取是否有效将在一定程度上影响关键词提取质量的好坏。文章、词汇链和《同义词词林》语义词典中提供了大量的特征信息,与其他方法只采用 TFIDF、位置的统计特征不同,在词语权重计算时,考虑了将词频、重要区域、词跨度、首次出现位置与词汇链强度特征等相关信息相结合,提出了一种改进的权重计算式(17),具体特征信息及描述如表 1 所列。

表 1 词语特征描述

类别	特征	简要描述
词频	TFIDF	表示某个词在文档中出现的频率,并且和文档集中出现该词的文档数作比较。
位置	重要区域 Area	表示出现在文档标题、摘要和章节标题等重要区域的词汇比其他词汇是关键词的可能性要大。
	词跨度 W_span	表示词在文章中出现的范围,跨度越大说明该词越能反映文章主题。
	首次出现位置 W_pos	表示文档中该词首次出现位置之前的词汇量占文档所有词汇总量的比率,取值在 0—1。一般出现在文档开头和结尾位置的词成为关键词的可能性比较大。
词汇链强度	词汇链长度 L_length	链中包含的词的数目,值越大说明该链越能表达文章的主题。
	词汇链跨度 L_span	指该链中的词最后出现位置与最早出现位置的距离,反映词汇链所表示的主题在文中被覆盖的范围,值越大则反映的主题在文中越重要。
	覆盖句子密度 L_sent	该特征表示链覆盖的密度,其值为包含链中任意词的句子数。
	关联度 L_rel	词 w_i 与链中其余词之间的相关度的均值,表示该词的语义强度,值越大表明该词与周围环境关联越紧密。

表 1 中经典的词语 TFIDF 计算公式为

$$TFIDF_{(w_i,d)}=\frac{tf(w_i,d)\times\log_2(\frac{N}{n_{w_i}})}{\sqrt{\sum_{w_i\in d}(tf(w_i,d)\times\log_2(\frac{N}{n_{w_i}}))^2}} \tag{11}$$

式中,$TFIDF_{(w_i,d)}$ 表示词 w_i 在当前文档 d 的 TFIDF 值,$tf_{(w_i,d)}$ 为词 w_i 在文档 d 中的出现次数;N 为语料库中总的文档数;nw_i 为语料库中含有词 w_i 的文档数。

词汇链跨度 L_span 计算公式为

$$L\_span=\frac{Length_{jl}-Length_{jf}}{Length} \tag{12}$$

式中,$Length_{jf}$ 为链 L_j 中的词在文档中最早出现之前的词汇数;$Length_{jl}$ 为链 L_j 中的词在文档中最后出现之前的词汇数;$Length$ 为文档中词汇总数。

词汇链关联度 L_rel 的计算公式为

$$L\_rel=\begin{cases}\dfrac{\sum_{i=1,j\neq i}^{L\_length}rel(w_i,w_j)}{L\_length-1}, & \text{if } L\_length>1\\ 0, & \text{if } L\_length=1\end{cases} \tag{13}$$

式中,L_length 表示该词汇链的长度,即词汇链所包含的词汇数目。

根据上述词汇链强度的 4 个特征信息,对构建后的每条词汇链进行链强度计算,即

$$T_i=\frac{(c_1\cdot L\_length+c_2\cdot L\_span+c_3\cdot L\_sent+c_4\cdot L\_rel)}{\sqrt{(L\_length^2+L\_span^2+L\_sent^2+L\_rel^2)}} \tag{14}$$

式中,T_i 为词汇链的强度;L_length 为词汇链包含的词数目;L_sent 为包含链中任意词的句子数。链跨度 L_span 和关联度 L_rel 分别按式(11),式(12)计算,c_1,c_2,c_3,c_4 为调节因子,分别取为 0.1,0.3,0.2,0.4,$c_1+c_2+c_3+c_4=1$;式中分母作归一化处理。

每个文本表示成 $T=\{T_1,T_2,\cdots,T_n\}$,其中 T_i 表示各个词汇链的权值。权值越大,表示链强度越高,即词汇链表达主题信息的能力越强;反之,权值越小,表达主题越弱。根据预设的阈值,从中选取最强的几个词汇链来共同表达文本,其中包含未登录词所在的词汇链。

词跨度 W_span 的计算公式为

$$W\_span=\frac{Length_l-Length_f}{Length} \tag{15}$$

首次出现位置 W_pos 的计算公式为

$$W\_pos=|\frac{Length_f}{Length}-0.5| \tag{16}$$

式中,$Length_f$ 为 w_i 所在文档中该词汇首次出现之前的词汇数;$Length_l$ 为 w_i 所在文档中该词汇最后一次出现之前的词汇数;$Length$ 为文档中词汇总数。

在综合考虑词汇在文档中的频率信息、区域位置信息和所处词汇链的强度等特征信息后,提出了词汇权重计算方法,即

$$Weight_i=\alpha\times TFIDF_{(w_i,d)}+\beta\times Area_i+\gamma\times T_i+\mu\times W\_span+\eta\times W\_pos \tag{17}$$

式中,$Weight_i$ 为词汇 w_i 的权值;$TFIDF_{(w_i,d)}$,T_i,W_span,W_pos 分别按式(10),式(13)一式(15)计算;$Area_i$ 为 w_i 所处文档区域的权重,如果 w_i 出现在文档标题,$Area_i=3$,如果出现在摘要中,$Area_i=2$,如果出现在章节标题,$Area_i=1.5$,否则,$Area_i=0.5$;T_i 为 w_i 所在词汇链的权重;α,β,γ,μ 和 η 是各属性的调节因子。

至此,对词汇链 T_i 中的所有词汇进行权重计算后,可用 $T_i=\{t_{i1},t_{i2},\cdots,t_{im}\}$ 来表示,其中 t_{ij} 表示构成词汇链 L_i 的词汇 w_j 的权值信息。对词汇链中所有词汇的权值进行计算,并按降序排序,根据预定的关键词需求个数,依次选取权值最大的词汇作为关键词,最终实现从中文文本中提取出能较全面较准确表达文本主题信息的关键词。

4 实验结果与分析

4.1 实验数据

为了对基于 KESELC 的关键词提取方法进行评价,采用较有影响力的复旦大学计算机信息与技术系国际数据库中心自然语言处理小组提供的含有 20 个类别的中文语料库[19],选取了计算机、教育、航空航天、历史、艺术、农业、环境、经济、政治和体育 10 个领域中的文本作为实验数据,每个领域随机选择 30 篇赋予关键词的长度适中文本作为语料集。首先随机选取其中的 150 篇文档作为学习样本来确定第 3 节中阈值 σ 与调节因子 α,β,γ,μ 和 η,通过反复实验,分别取为 0.5,1.2,1.0,1.5,1.0,1.0。

4.2 实验评估标准

对关键词提取算法的评估方法是将算法提取出来的关键词与人工赋予的关键词作匹配比较,实验时以文档作者提供的关键词作为标准答案,同时词法上的匹配方法由于没有考虑语义关系,因此不能充分表达实际的匹配效果。如算法提取和人工赋予的关键词分别为“开心”和“高兴”,则该方法会认为匹配不成功。这显然不切实际,所以本文采用更为合理

的基于语义的评估方法，通过相似度计算，将同义的词语也认为匹配成功。

评价采用比较常用的查准率 Precision(简记为 P)、查全率 Recall(简记为 R)和平衡两者的综合指标 F-measure 对关键词提取算法进行评价。具体如公式分别为：

$$P=\frac{a}{b} \tag{18}$$

$$R=\frac{a}{c} \tag{19}$$

$$F\text{-}measure=\frac{2\cdot P\cdot R}{P+R} \tag{20}$$

式中，a 为正确匹配的关键词个数；b 为所有提取的关键词个数；c 为人工赋予的关键词个数。

4.3 实验结果分析与比较

为了验证 KESELC 算法特征选取的效果，首先采用 F-measure 综合指标进行评估，通过选取不同特征项来比较分析文本关键词提取的效果。由于人工赋予的关键词个数一般为 3～5 个，最多不大于 10 个，因此实验提取关键词个数控制为 3～10 之间。

将词频、区域位置、词汇链强度选取不同类特征划分组合为 3 组：(1)TFIDF 特征；(2)TFIDF＋位置特征集；(3)TFIDF＋位置＋词汇链强度特征集。KESELC 算法在 3 组不同特征选取组合上的关键词提取结果的对比如图 3 所示。

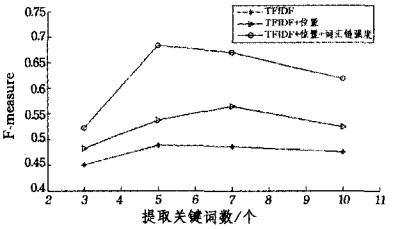

图 3 KESELC 算法的不同特征选取组合对比

由图 3 可以看出，与仅基于 TFIDF 特征的关键词提取相比，综合考虑位置特征信息、词汇链强度特征信息的关键词提取在 F-measure 综合评估上有显著的提高。基于 TFIDF 特征的关键词提取方法对出现频率较高的词汇均有所倾向。增加位置特征信息对出现于文本重要位置等分布特殊的词汇有所考虑。而增加词汇链强度特征信息弥补了词频位置等信息的不足，充分考虑了词汇间的语义信息，既考虑语义相似度又将语义相关度的语义扩展度和词汇链方法相结合，进一步有助于挖掘出频率不高但对表达文章主题具有重要贡献的词汇。

为了验证结合语义扩展度和词汇链的关键词提取算法 KESELC 的效果，采用查准率、查全率和 F-measure 上述 3 项指标进行评估，将本算法分别与比较常见且应用较广的基于 TFIDF 统计方法和同样为语义分析层面的传统的基于词汇链方法进行比较分析。实验结果如表 2 和图 4 所示。

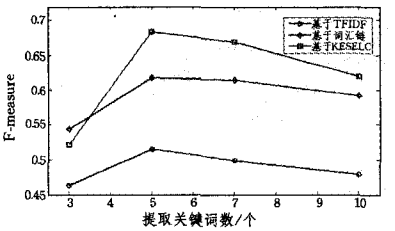

图 4 3 种算法的性能结果比较

表 2 不同关键词提取方法的实验结果比较

提取方法	提取关键词数/个	Precision	Recall	F-measure
基于 TFIDF	3	0.596	0.378	0.463
	5	0.577	0.465	0.515
	7	0.508	0.490	0.499
	10	0.426	0.547	0.479
基于词汇链	3	0.680	0.453	0.544
	5	0.657	0.581	0.617
	7	0.588	0.642	0.614
	10	0.503	0.719	0.592
基于 KESELC	3	0.711	0.412	0.522
	5	0.692	0.675	0.683
	7	0.634	0.709	0.669
	10	0.527	0.751	0.619

从实验结果中可以看出：当提取关键词个数小于 10 时，与仅基于统计的 TFIDF 关键词提取方法相比，基于 KESELC 的关键词提取方法在平均准确率、平均召回率和综合指标 F-measure 3 项指标上都有明显提高。基于 KESELC 方法所获得的平均性能总体上也要稍优于基于词汇链的关键词提取方法。当提取关键词个数为 3 时，基于 KESELC 的关键词提取方法性能稍低于基于词汇链方法，随着关键词提取个数的增加，基于 KESELC 方法的提取效果逐渐显现出来，并最后趋于稳定。同时可以发现，基于 KESELC 方法与其他两个方法的 F-measure 值均在关键词提取个数为 5 时获得最优值。

结果表明：文章一般被赋予的关键词个数为 5 个左右，自动提取关键词个数与人工赋予的关键词个数相一致时，可获得较优的提取效果。基于 TFIDF 关键词提取方法单纯从词频角度考虑关键词的重要性，提取的关键词可能存在一些与文章主题信息无关的高频词，从而降低了提取的质量。基于词汇链的方法深入到了语义一级，但还尚未很好解决同义词冗余表达、文章主题反映不够确切等问题。基于 KESELC 方法在《同义词词林》词典的基础上将语义扩展度和词汇链方法相结合，充分考虑词语间语义关系，采用共现率模型有效解决了多义词歧义分歧问题，同义词合并处理有效避免了同义词的冗余表达，对同义词比较多的文章来说，该方法提取关键词的效果更好，并且最终提取出的关键词很好地覆盖了文章的多个主题。同时可以发现，当提取关键词个数为 3 时，基于 KESELC 的关键词提取方法性能稍低于基于词汇链方法，主要原因为关键词个数较少时，结合语义相似度和语义相关度的语义扩展度方法不能充分发挥作用，随着关键词提取个数的增加，基于《同义词词林》的语义扩展度和特征集的优势发挥作用，有助于挖掘出表达文本主题的潜在信息，基于 KESELC 方法的提取效果有明显提高，并最后趋于稳定。

结束语 在充分利用《同义词词林》丰富语义知识的基础上，将语义扩展度和词汇链方法相结合，提出了一个完整的基于语义的关键词提取算法 KESELC。首先对预处理后的候选关键词进行词义消歧，然后根据《同义词词林》语义知识库的特点和统计信息，通过计算由词汇间语义相似度和相关度得出的语义扩展度来构建词汇链，综合考虑词汇所在词汇链的强度、词频、区域位置等特征信息，并在此基础上改进了权重计算来提取关键词。实验结果证明：通过语义角度分析，采用共现率模型有效解决了多义词歧义分歧问题，同义词合并处理避免了同义词的冗余表达，对同义词比较多的文章来说，

(下转第 291 页)

hierarchical binary decision tree approach [J]. Speech Communication, 2011, 53(9/10): 1162-1171

[9] Tecce J J. Psychology, physiological and experimental [Eyeblinks and psychological functions](6th ed)[M]// McGraw-Hill Yearbook of Science and Technology. NY: McGraw-Hill, 1992: 375-377

[10] Harrigan J A, O'connell D M. How do you look when feeling anxious? Facial displays of anxiety [J]. Personality and Individual Differences, 1996, 21(2): 205-212

[11] Matsuo T. Availability of eye blinks as index of cognitive anxiety on usability evaluation [J]. Japanese Journal of Ergonomics, 2004, 40(3): 148-154

[12] Patel M, Lal S, Kavanagh D, et al. Fatigue Detection Using Computer Vision [J]. International Journal of Electronics and Telecommunications, 2010, 56(4): 457-461

[13] Chau M, Betke M. Real time eye tracking and blink detection with USB cameras [D]. Boston, MA: Boston University, 2005: 2215

[14] Wu J, Trived M M. Simultaneous eye tracking and blink detection with interactive particle filters [J]. EURASIP Journal on Advances in Signal Processing, 2008, 2008: 1-17

[15] Horng W-B, Chen C-Y, Chang Y, et al. Driver fatigue detection based on eye tracking and dynamic template matching[C]// Proceedings of the IEEE International Conference on Networking, Sensing and Control. 2004

[16] Pearl J. Probabilistic reasoning in intelligent systems: Networks of plausible inference[M]. US: Morgan Kaufmann, 1988

[17] Csikszentmihalyi M. Flow: The psychology of optimal experience [J]. New York: Harper Perennial Modern Classics, 1990: 314

[18] Gagne R M, Wager W W, Golas K C, et al. Principles of instructional design [J]. Performance Improvement, 2005, 44(2): 44-46

[19] Lauritzen S L, Spiegelhalter D J. Local computations with probabilities on graphical structures and their application to expert systems [J]. Journal of the Royal Statistical Society Series B (Methodological), 1988, 50(2): 157-224

[20] Lauritzen S L. The EM algorithm for graphical association models with missing data [J]. Computational Statistics & Data Analysis, 1995, 19(2): 191-201

(上接第 269 页)

该方法提取关键词的效果更好，并且最终提取出的关键词很好地覆盖了文章的多个主题。基于 KESELC 的关键词提取方法在准确率、召回率和综合指标 F-measure 方面与其他方法相比平均性能上均有所提高。同时可以发现，除了分词等影响之外，《同义词词林》的规模也制约了对关键词的提取。将来进一步开展的工作方向：根据未登录词具有最新热点信息可能性较大和组合词富含信息量较多的情况，需要改善分词算法的分词与未登录词、组合词的识别能力；另一方面需要进一步扩充语义知识库，尤其是专业术语领域的词汇，进一步发掘出非高频而重要性高的关键词语。

参 考 文 献

[1] Bao Hong, Deng Zhen. An extended keyword extraction method [C]// Proceedings of the 2012 International Conference on Applied Physics and Industrial Engineering. USA: Elsevier, 2012: 1120-1127

[2] 李霞，李战怀，张利军，等. MXDR：一种基于关键字的 XML 多文档分布式检索方法[J]. 计算机科学，2011，38(10)：152-156

[3] 郑斐然，苗夺谦，张志飞，等. 一种中文微博新闻话题检测的方法[J]. 计算机科学，2012，39(1)：138-141

[4] G'abor B, Rich'ard F. SZTERGAK: Feature engineering for keyphrase extraction[C]// Proceedings of the 5th International Workshop on Semantic Evaluation. Sweden: ACM, 2010: 186-189

[5] Witten I H, Paynter G W, Frank E, et al. KEA: Practical automatic keyphrase extraction[C]// Proceedings of the 4th ACM Conference on Digital Libraries. Berkeley, California, US: ACM, 1999: 254-256

[6] Lopez P, Romary L. HUMB: automatic key term extraction from scientific articles in GROBID[C]// Proceedings of the 5th International Workshop on Semantic Evaluation. Uppsala, Sweden: ACM, 2010: 248-251

[7] 苏金树，张博锋，徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报，2006，17(9)：1848-1859

[8] 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

[9] Meng Wen-chao, Liu Lian-chen, Dai Ting. A modified approach to keyword extraction based on word-similarity[C]// Proceedings of the 2009 IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS). Shanghai, China: IEEE, 2009: 388-392

[10] Li Gang, Dai Qiang-bin, Wei Quan. A new approach to compute semantic relevance of Chinese words[C]// Proceedings of the 2010 IEEE International Conference on Artificial Intelligence and Education (ICAIE). Wuhan, China: IEEE, 2010: 610-613

[11] 聂卉，龙朝辉. 结合语义相似度与相关度的概念扩展[J]. 情报学报，2007，26(5)：728-732

[12] LI Xing-hua, WU Xin-dong, HU Xue-gang, et al. Keyword extraction based on lexical chains and word co-occurrence for Chinese news Web pages[C]// Proceedings of the 2008 IEEE International Conference on Data Mining Workshops. Pisa, Italy: IEEE, 2008: 744-751

[13] 梅家驹，竺一鸣，高蕴琦，等. 同义词词林[M]. 上海：上海辞书出版社，1993：106-108

[14] 陆洋. 基于语义分析的文本挖掘研究[D]. 杭州：浙江工业大学，2011

[15] Institute of Computing Technology, Chinese Academy of Sciences. ICTCLAS [EB/OL]. http://ictclas.org/index.html, 2012-04-01

[16] 田久乐，赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报：信息科学版，2010，28(6)：603-608

[17] Satanjeev B, Ted P. Extended gloss overlaps as a measure of semantic relatedness[C]// Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence. Acapulco, Mexico: Aminer, 2003: 805-810

[18] Jane M, Graeme H. Lexical cohesion computed by thesaural relations as an indicator of the structure of text[J]. Computational Linguistics, 1991, 17(1): 21-48

[19] Li Rong-lu. Fudan university text corpus [DB/OL]. http://www.nlp.org.cn/docs/doclist.php? cat_id=16&type=15, 2012-04-01