

基于词汇链的预案主题抽取方法研究

裘江南¹ 罗志成² 王延章¹

(1. 大连理工大学管理学院, 大连, 116024, 2. 武汉大学信息管理学院, 武汉, 430072)

摘要: 本文针对应急预案自动主题抽取的需求, 致力于词汇语义相关度的计算, 构建了一个基于词汇链算法且符合人的主观感受的主题抽取模型。模型根据应急预案文本的特点, 运用了若干自然语言处理技术, 改进了原始的词汇链生成算法, 提出了一种多因素词语权重算法。最终, 通过与人工主题词抽取的实验结果相比较, 该主题提取模型在查全率和查准率上都取得了较好的效果。

关键词 主题抽取, 词汇链, 语义相关, 应急预案

Research on Semantic Relatedness Based Subjects Extraction from Emergency Plans

Qiu Jiangnan¹ Luo Zhicheng² Wang Yanzhang¹

(1. School of Management of Dalian University of Technology, Dalian 116024,

2. School of Information Management, Wuhan University, Wuhan 430072)

Abstract: The paper aimed at the requirement of the automatic extraction of subject from the emergency plans, took up with the measures of lexical semantic relatedness, and has constructed a subject extraction model based on the lexical chain algorithm which accords with human's subjective feeling. According to the characteristics of the emergence plans text and the needs of the project, the model used a number of natural language processing methods, improved the original chain generating algorithm, and brought forward a weight algorithm base on multi-factors. Finally, an experimental was carried out which compared the human subject extraction results to our system result, and the recall and the precision showed that our model do a good job.

Keywords Subject Extraction, Lexical Chain, Semantic Relatedness, Emergency Plans

1 引言

应急管理的过程中, 预案是应急决策和指挥者依法处置的法律依据, 应急决策相关主题知识段落可能是一篇预案文本、一篇预案文本的一部分或几篇预案文本各部分的集合。随着中央政府、各部委、各地方政府发布的预案数量与日剧增。因此, 能够快速、准确、全面地从众多预案中提取相关文本信息是应急辅助决策信息系统的主要功能, 而其中的核心基础是实现从大量的文本中抽取出用户相关的主题性知识段落。传统的全文检索方法可以提供文本段落的定位功能, 但是全文检索的核心是关键字的机械式匹配, 所以经常出现检索不全、答非所问的结果^[1], 因而传统全文检索方式难以满足应急管理的需要。

本文针对预案的文本特点, 采用预案文本结构化和文本章节主题抽取的信息组织方法, 为快速准确的知识定位和检索打下基础。有关中文主题抽取和标引方面学术界已作了较多的研究, 具体包括: 1) 王永成等人^[2]建立了中文文献主题自动标引系统, 提出了采用实词的相对频率、特征词, 并结合词形聚类的主题关键词加权标引算法。2) 李素建等^[3]提出了利

作者简介: 裘江南, 男, 1968年生, 副教授, 博士研究生, 研究方向为电子政务、知识管理。Email: qiu_jn@tom.com。罗志成, 男, 1984年生, 硕士研究生, 研究方向为信息检索。Email: luozhicheng.dut@gmail.com。王延章, 男, 1952年生, 教授, 研究方向为电子政务、决策支持系统。

用最大熵模型进行关键词自动标引的方法，由于特征参数估计的误差，导致最终查全率和查准率都不理想。3) 索红光等人^[4]提出一种基于词汇链的主题抽取方法，并取得了较好的效果，但该项研究由于采用刘群^[5]的没有提供标准接口的《知网》相似度计算软件包，导致词汇链算法的使用受到很多限制。

基于词汇链的主题抽取方法是近年来提出的一种新方法，而最初引入词汇链的主要目的是用于分析文本的结构。应急预案是一种较为规范和结构性较好的文本，因此，对预案的主题抽取和标引可采用基于词汇链的方法。

2 系统分析

2.1 词汇链算法分析

词汇链算法是Morris和Hirst于1991提出的^[6]，其中词汇链是指一个主题下的一系列相关的词共同组成的词系列。词汇链算法的原理是：在文章中描述某个主题的文本块内，使用的词语应该是相关的，这些相关词语构成一条词汇链。所以，词汇链可以视作一个语言片段的标志性主题词语链，不同的词汇链对应了不同的语言片段。因此，一旦词汇链确定，那么文章的结构也就确定了。

Morris和Hirst最初使用词汇链的目的是用于文本分割，即分析文本的结构。其基本想法是：由于词汇链是一系列相关的词所组成的，这些词表达的是同一件事情或意思，找到这些链就得到了文本的结构。后来这一基本想法在很多方面得到了应用，比如文本检索、信息抽取、检查文本的用词不当等。

目前国内对于词汇链的研究较少。最早的是刘素红^[7]等人对词汇链算法的介绍，之后有尤文建^[8]基于词汇链构建文本过滤模型。另外，索红光^[4]利用改进的词汇链算法和刘群开发的《知网》相似度计算软件包，提出了一种关键词抽取方法，但是实验结果的查全率和查准率都比较低。陈燕敏等人^[9-10]将词汇链算法应用于自动文本摘要，实验结果表明，他们算法的查全率和查准率都比较高。

基于上述分析，本文考虑首先将一篇文本中的词汇按照它们的词义相关度构建多个词汇链，然后按照一定的规则从中挑选出能够代表文本主题的关键词。其中，度量词汇语义相关性是生成词汇链的基础，根据文献^[11]的实验结果，本研究在词汇链的生成中采用基于语义词典的语义相关度量方法。

2.2 应急预案特点分析

应急预案的文本具有比较强的结构化特性，及明显的章节编号，这些章节编号有助于提取出文本的篇章结构。因此，应急预案的是一种“准规范文本”。其具体的特点总结如下：

- 1) 文本中使用一些可辨识的符号，如第三章、第2节等等。一个编号所统领的文字区域成为一个章节；
- 2) 文本中一个章节包含多个子章节，子章节用不同的符号来标记；
- 3) 文本中存在不可再分的子章节，可称之为原子章节；
- 4) 文本中一个章节中除了子章节之外，可以含有其他段落，这些段落中不包含章节符号，可称之为该章节的附属段落；
- 5) 在同一文本中，同一个章节中一般只包含附属段落或者子章节；
- 6) 文本中的用语规范严谨，章节标题一般概况了该章节的主要内容。

3 算法流程设计

根据前文的分析，设计了主题抽取的流程如图 1 所示。下文对图中各个操作进行详细介绍。

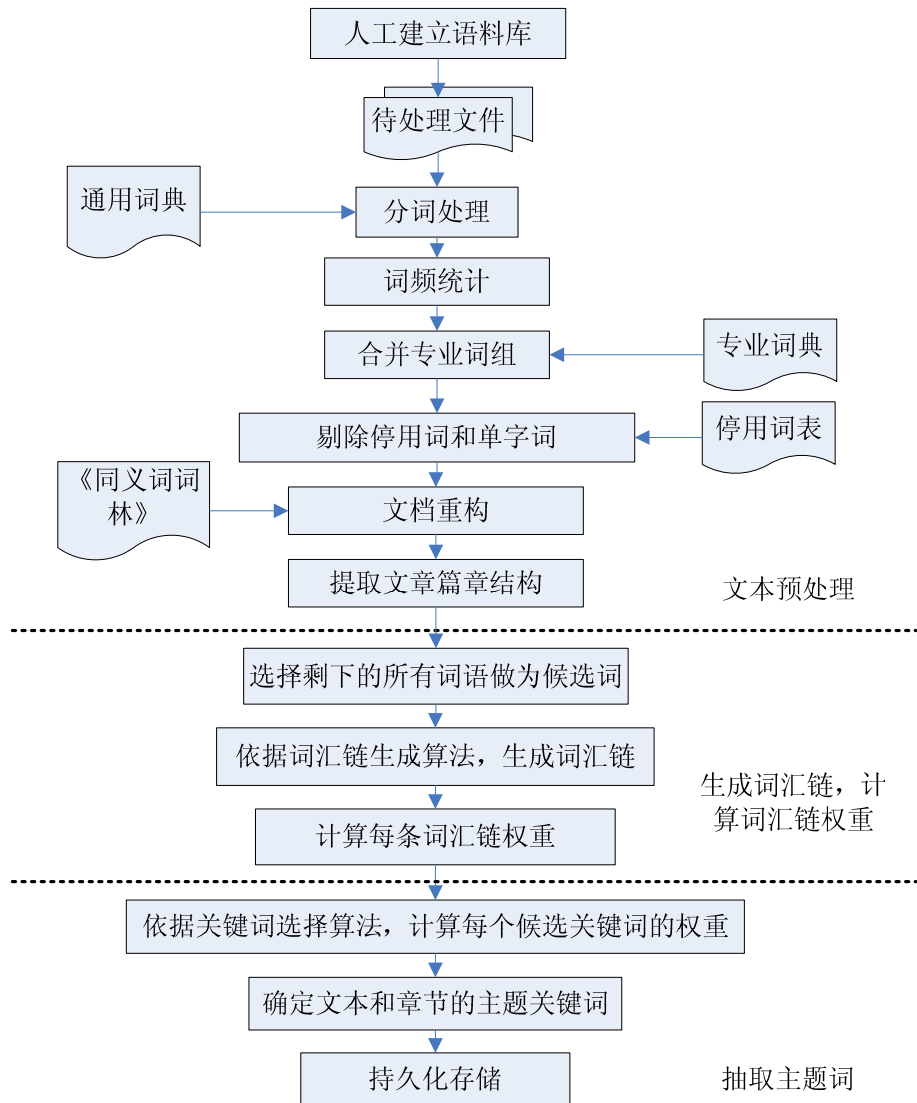


图 1 主题抽取流程图

3.1 文本预处理

分词是中文、日文等亚洲语言的信息检索中遇到的特殊问题，目前中文分词技术已日益完善，本研究采用中科院计算技术研究所发布的 ICTCLAS-3 软件进行分词。同时，根据停用词表剔除停用词，为后面词汇链生成做准备。

因为文本的主题词有的是专业词语，而得到的分词结果却将这些词语分割开，所以模型中根据专业词典对分词后的这部分专业词语重新合并。之后，根据事先制订的停用词表剔除停用词。同时，考虑到汉语当中的单字词所含信息量很少，所以也把所有单字词剔除。

本研究需要抽取的是篇章的主题词，而一个篇章的数据量相对比较小。这样难免出现数据稀疏的问题，因而本研究在统计词频的之前进行文本重构。通过借鉴张敏等人提出的文本重构的方法^[12]，将文本中的主题词的下位词替换为上位主题词，以增加上位主题词的词频。同时，对于同一个集合中的词语，将低频词语替换为高频词语，以进一步增加高频词语的频率。

3.2 提取文章篇章结构

因为应急管理当中的检索需要定位到相关的主题性知识段落,所以文本结构化是应急管理系统的必要工作。通过 2.2 节中的分析,可知应急预案大多都符合准规范文本。所以可以通过文本的特征直接提取文本的篇章结构,从而实现文本结构化。而文本的章节往往表示为若干个主题知识,本研究在提取文本主题结构的基础上,对章节应用词汇链方法来抽取主题词。

在具体的提取方法上,本文借鉴单永明提出的汉语文本形式结构分析及其标引算法^[13],并进一步细化的算法流程如图 2 所示。

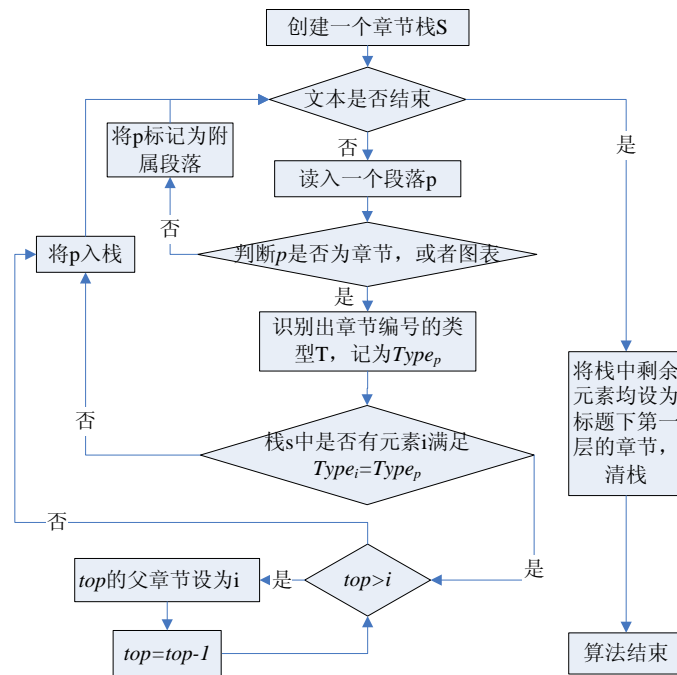


图 2 文本篇章结构提取算法

3.3 生成词汇链

本研究采用哈尔滨工业大学信息检索实验室提供的《同义词词林》作为语义词典^[14],并通过改进 Morris 和 Hirst 的方法,提出了词汇链生成算法,其流程如图 3 所示。在构建词汇链时,需要根据词义相关度的阈值来判定。因此,算法中相关度阈值的设置对构建词汇链有非常大的影响。当阈值过高时,每个链中词的数目会很少,链的数目很多;而当阈值过低,链的数目减少,每个链的词数很多。我们在实验中发现随着文本长度的增加,即使是一个较大的阈值,也会出现词汇链过长的问题。限制词汇链的数量只会使得词汇链越来越长,最终词汇链中的词语权重趋同。最终,本研究采用规定词汇链的最大长度的方法解决此问题。

由于本研究是从章节中抽取主题,章节中相应词语的词频都较低,因此,本研究与索红光的方法^[4]不同,并不按照词频大小作为选择候选词的依据,而是把文本预处理之后的所有词语作为候选词。另外,根据研究中开展的词汇对相关度人工判断的实验结果,取出人工判断相关度最高的 10 对词,对这些词对的系统计算结果求平均值,微调之后作为词汇链相关度的阈值。

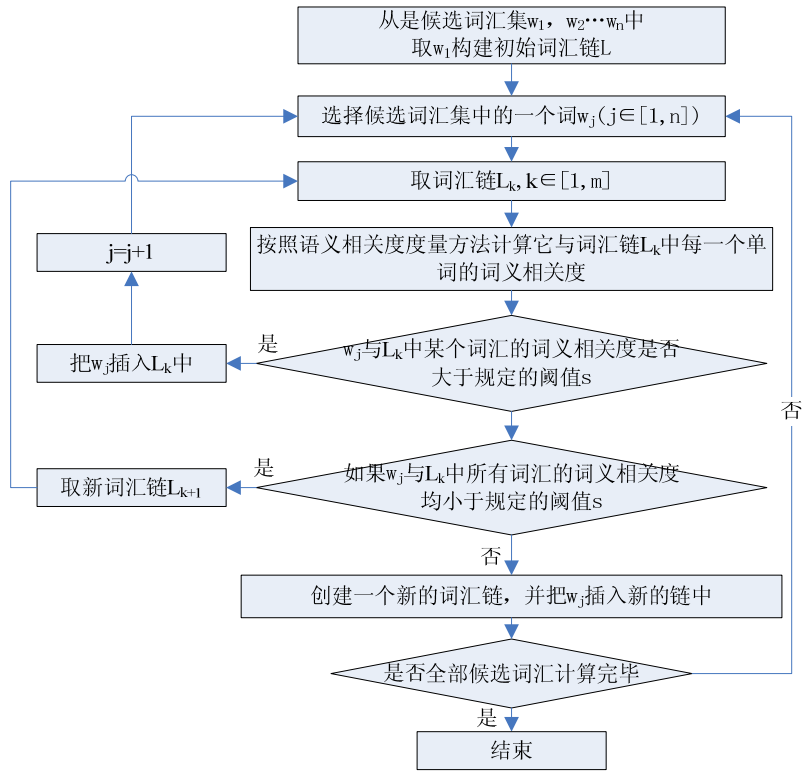


图3 词汇链生成算法流程图

3.4 抽取主题词

抽取主题词，即找出权重最大的词语。本研究在综合借鉴郑家恒等的词频及区域函数^[15]和索红光的算法基础上，考虑词频、区域因子和词汇链权重等三个因素，提出一种基于词汇链的关键词权值函数如公式(1)所示：

$$weight_i = a \cdot freq_i + b \cdot loc_i + c \cdot chainWeight_i \quad (1)$$

其中， $weight_i$ 表示词汇*i*的权值；

$freq_i$ 表示词汇*i*的词频因子。词频与词语的重要性呈正相关关系，这符合人们的主观认识，即一个词语在文章中出现的次数越多，那么它往往越重要。当然，那些频率极高的词汇应该被剔除，因为这些词语出现在每一篇文章中，那么它们的代表性就大大降低。但是在应急管理的实际项目当中，因为是在一个章节甚至仅仅一个段落中比较词语权重，词频往往都较低，所以仅仅考虑词频是不够的。

公式(1)的 loc_i 表示词汇*i*的区域因子，依据预案文本的特点及郑家恒在文献^[15]提出的思想，本文将区域因子权重函数按标题、主题句、普通文本等三部分定义，如公式(2)所示。

$$loc_i = \begin{cases} 5 & \text{如果词语在文章标题中} \\ 2 \times \frac{tFreq}{tFreq + 1} & \text{如果词语在主题句中} \\ 1 & \text{如果词语在普通文本区域} \end{cases} \quad (2)$$

(tFreq表示该词语在主题句中出现的次数)

公式(1)中的 $chainWeight_i$ 表示词汇 i 所在词汇链的权重，可根据公式(3)计算。

$$chainWeight = length + strongWordNum \quad (3)$$

公式(3)中的 $length$ 为词汇链的长度， $strongWordNum$ 是指词汇链中包含在标题中和主题句中的词语的个数。

公式(1)中的 a 、 b 和 c 是 $freq_i$ 、 loc_i 和 $chainWeight_i$ 的调节因子。根据本文最终的实验经验， b 应该取一个相对较大值，因为相对于区域因子的结果值而言，词频和词汇链权重的结果值都比较大。在本文第 4 节多因素重要性比较的实验中，在抽取文档主题词的时候， $a=1, b=4, c=0$ ；在抽取章节主题词的时候， $a=0, b=2, c=1$ 。

4 主题词抽取效果分析

下面将文中提出的主题词自动抽取方法与人工抽取方法的试验结果作对比分析，来评价本文所提出方法的效果。实验的测试集是从应急预案中随机选取 10 篇预案，这些应急预案的特征如表 1 所示。三个受试者各自独立地从整篇文档和文档的各个章节中抽取出若干个主题词，具体主题词数量由个人视文本块长度而定，从而构建了 1983 个“理想”章节，30 篇“理想”文档。之后，应用本文提出的主题词抽取方法来抽取主题词，与人工抽取的“理想”主题词作比较，通过计算平均查准率（Precision），查全率（Recall）和调和平均值F值来评价系统抽取主题词的质量^[10]。

表 1 实验所采用的应急预案的特征

编号	文本名称	章节数目	字数
1	国家安全生产事故灾难应急预案	52	6607
2	国家处置城市地铁事故灾难应急预案	78	7440
3	国家处置电网大面积停电事件应急预案	40	4915
4	国家处置民用航空器飞行事故应急预案	50	7117
5	国家突发重大动物疫情应急预案	58	8260
6	国家重大食品安全事故应急预案	53	7224
7	国家突发环境事件应急预案	73	10045
8	国家通信保障应急预案	44	4833
9	国家防汛抗旱应急预案	147	14944
10	国家地震应急预案	66	9145

在此实验条件下，我们分别计算只考虑词频、只考虑区域因子、只考虑词汇链权重（采用Resnik的度量方法^[16]，阈值为 6.86）、同时考虑词频和区域因子、同时考虑词频和词汇链权重、同时考虑词汇链权重和区域因子、同时考虑词频和词汇链权重和区域因子等 7 种情况下的主题词抽取效果。

表2 词频和词汇链权重和区域因子重要性分析

方法	章节			整篇文本		
	查全率	查准率	F 值	查全率	查准率	F 值
词频	0.461	0.514	0.486	0.913	0.520	0.663
区域因子	0.782	0.682	0.729	0.261	0.260	0.260
词汇链	0.252	0.293	0.271	0.174	0.180	0.177
词频-区域因子	0.740	0.686	0.712	0.913	0.520	0.663
词汇链-区域因子	0.779	0.680	0.726	0.565	0.600	0.582
词汇链-词频	0.470	0.521	0.494	0.478	0.540	0.507
词汇链-词频-区域因子	0.739	0.683	0.710	0.565	0.580	0.573

从表2中可以看出,区域因子在章节主题词的抽取中起了最重要的作用。这实际上符合本文之前分析的应急预案文本特点,即应急预案用语规范严谨,章节标题都概况了该章节的主要内容。但是如果考虑区域因子,那么在抽取整篇文本的主题词的时候,查全率和查准率都很低。

词频在整篇文本主题词抽取上效果最好,查全率达到了最高值91.3%。其中一部分是因为文本主题词数据量比较小,另外也反应了词频对于主题抽取的重要性。但是在章节主题词抽取之中,仅仅考虑一个词语在该章节中的出现频率是不够的。因为章节中的词频都偏低,并且某些重要词语如章节标题中的词语可能仅仅出现一次。

如果说仅仅考虑词汇链的权重的话,那么效果是最差的。但是在和区域因子结合之后,章节主题词抽取的效果是很好的,文本主题词抽取的效果相对较好。这和我们最初的预期非常符合,即在章节这种词语量很小的处理中,在抽取主题词的时候,仅仅考虑词频是不合理的,而应该把较为相关的词语也考虑进来。而生成词汇链之后,考虑词汇链权重,就相当于是一种扩展的词频统计结果,而不必再考虑单纯的词频。

通过上述分析和综合考虑,本研究拟采用词汇链-区域算法,即在计算词语权重的时候,考虑词语的词汇链权重和区域因子。最终主题词抽取的效果如表3。其中词汇链算法中相关度量方法采用Resnik的方法^[1],阈值为6.86。

表3 主题抽取效果表

章节			整篇文本		
查全率	查准率	F 值	查全率	查准率	F 值
0.779	0.680	0.726	0.565	0.600	0.582

5 结论

本文提出了一个基于词汇链的应急预案的主题提取模型。模型中根据应急预案文本的特点和项目的需要,运用了若干自然语言处理技术,改进了原始的词汇链生成算法,提出了多因素词语权重算法。并建立了一个主题词自动抽取系统。通过与人工主题词抽取方法的实验结果相比较,该主题提取模型在查全率和查准率上都取得了较好的效果。进一步的工作中,将调整系统中用到的众多参数和经验值,以求获得一种最优的主题抽取方案,进一步提高系统的查全率、查准率和执行效率。

参考文献

- [1] 李蕾,王楠,钟义信等.基于语义网络的概念检索研究与实现.情报学报,2000,19(5):525-531.

- [2] 王永成, 顾晓明, 王丽霞. 中文文献主题的自动标引. 情报学报, 1998, 17(3):219-225.
- [3] 李素建, 王厚峰, 俞士汶等. 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.
- [4] 索红光, 刘玉树, 曹淑英. 一种基于词汇链的关键词抽取方法. 中文信息学报, 2006, 20(6):25-30.
- [5] 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. 2002, 7(2):59-76.
- [6] Morris J, Hirst G. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 1991, 17(1):21-48.
- [7] 刘素红, 刘传汉, 王永成. 动态词链算法. 计算机工程, 2003, 29(20):80-81.
- [8] 尤文建, 李绍滋, 李堂秋. 基于词汇链的文本过滤模型. 计算机应用研究, 2003, (9):32-35.
- [9] Chen Yanmin, Liu Bingquan, Wang Xiaolong. Automatic Text Summarization Based On Textual Cohesion. *Journal of Electronics (China)*, 2007, 24(3):338-346.
- [10] 陈燕敏, 王晓龙, 刘秉权等. 多知识源融合的自动摘要系统研究与实现. 高技术通讯, 2006, 16(4):337-341.
- [11] Mohammad S, Hirst G. Distributional Measures of Concept-Distance: A Task-oriented Evaluation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, Sydney, 2006:35-43.
- [12] 张敏, 宋睿华, 马少平. 基于语义关系查询扩展的文档重构方法. 计算机学报, 2004, 27(10):1395-1401.
- [13] 单永明. 汉语文本形式结构分析及其标引算法. 中文信息学报, 2001, 16(2):14-19.
- [14] 《同义词词林》扩展版. <http://www.ir-lab.org/>
- [15] 郑家恒, 卢娇丽. 关键词抽取方法的研究. 计算机工程, 2005, 31(18):194-196.
- [16] Resnik, P. Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, Canada, 1995.