朝鲜语信息检索索引方法研究

金光赫[1,2] 王兴伟[1] 蒋定德[1]

(东北大学信息科学与工程学院 沈阳 110819)[1] (金策工业综合大学应用程序学院 平壤)[2]

摘 要 基于朝鲜语信息检索系统的深入分析,研究提高朝鲜语信息检索性能的索引问题。通过剖析名词单位索引法、单位词素索引法、n-gram单位索引法、单位语句索引法等经典索引法的优缺点,以试验分析找出对索引性能有重要影响的关键要素,深入阐述朝鲜语的30个非用词、索引方式与朝鲜语的特征,从而提出一种新的将每种索引方法特征融于一体的朝鲜语信息检索索引方法。仿真实验表明,所提出的新方法具有更好的性能。

关键词 朝鲜语,词素分析,索引法,n-gram方法,非用词

中图法分类号 TP393 **文献标识码** A

Study on Indexing Method for Korean Information Search

JIN Guang-he[1,2] WANG Xing-wei[1] JIANG Ding-de[1]

(School of Information Science & Engineering, Northeastern University, Shenyang 110819, China)[1]

(School of Application Program, Kim Chaek University of Technology, Pyongyang, DPR Korea)[2]

Abstract Based on the sufficient analysis of the Korean information search system, this paper investigated the indexing method to improve the search performance. After the advantage and shortcoming of the typical indexing methods such as the noun unit indexing, the morphological analysis indexing, the n-gram unit indexing, the word segmentation unit indexing and so on, were analyzed in detail, the key factor impacting significantly on the search performance was found by trial and error. At the same time, thirty stop words in Korean, indexing way used to search, and its characteristics were illustrated. Finally, a new indexing method for Korean information search was proposed by taking advantage of every indexing method. Simulation results show that new method proposed holds the significant performance improvement and is promising.

Keywords Korean, Morphological analysis, Indexing method, N-gram method, Stop word

1 引言

随着网络技术的快速发展和网络业务的指数式增长,各种各样的信息以文件形式存储和共享,产生越来越多的海量信息。因此,如何有效地检索以文件形式存储的海量信息变得尤为重要。信息检索系统是把使用者所需要的信息收集分类后形成便于检索的形式,并保存为文件。信息检索时,在保存的文件中迅速找出对使用者有用的文件,根据信息检索要求提供相应的信息。而索引是对各种信息进行分析,提取其主要内容,并把它作为各种信息的摘要。使用索引可以缩小检索者检索的单词或检索的数量,从而缩小检索范围。

传统的信息索引方法主要是索引者或专业人士根据自身所拥有的知识对信息进行分析、归类,将一些关键词添加到索引库中。但是,在信息知识爆炸的今天,具有专业知识和索引经验的索引者却不能满足信息日益增长的需要,其结果导致信息检索面对巨大挑战,而海量信息分析、检索所消耗的巨大费用也是导致信息检索困难的另一重要原因。为了解决此类问题,出现了使用电脑符号进行自动分类的检索系统,这种自动索引方法是用电脑分析已有的文件后提取能代表该文献的单词或短句并把它作为索引词。

一般地,英语用空格来区分单词,提取单词后可以使用原型复原(stemming)的方法[24,25]对文件进行复原。但朝鲜语单词由几个单词组成,对索引词的提取难度大。从现有的朝鲜语信息检索系统来看,大多数为名词单位索引[11,12]、n-gram单位的索引[8,14-16,25-29]、词素单位索引[9,10]等。名词单位索引法用空格来区分语句,在识别语句后用最大匹配的方法(longest match strategy)提取了非索引分段以及非用语索引。n-gram索引法以空格提取区分语句,然后去掉非用语,用最大匹配方法对非索引分段语分段。词素单位索引法提取最小意义名词作为索引,其方法包括词素分析、去掉含糊性词语、提取单一名词和去掉非用语等4种形式。

本文通过分析、比较语句单位索引法、词素单位索引法、

到稿日期:2010-06-13 返修日期:2010-12-26 本文受国家自然科学基金(70671020,70931001,60802023),国家科技支撑计划(2008BAH37B03,2008BAH37B07),高等学校博士学科点专项科研基金(20070145017),中央高校基本科研业务费专项资金(N090504003,N090504006)资助。

金光赫(1978—),男,博士生,主要研究方向为朝鲜自然语言处理、检索、数据挖掘,E-mail:kghmsgfuture.0904@gmail.com;王兴伟(1968—),男,博士,教授,博士生导师,主要研究方向为下一代网络,E-mail:wangxw@mailneu.edu.cn(通信作者)。

n-gram单位索引法和名词单位索引法，剖析各种索引方法的优缺点，通过试验分析方法，提取对索引性能有重要影响的关键要素。同时，针对朝鲜语的语言特点，深入阐述朝鲜语的30个非用词和索引方式，从而在结合各种信息索引方法的基础上，提出一种新的朝鲜语信息检索索引方法。

2 索引词与非用语

索引词(indexing term)是为了检索文件内容而附加的，在检索方式上是以提问语(querying term)形式提出的用语，而不能以提问语形式提出的用语是非用词。因此，文件中最理想的索引词集合是检索此文件时可能成为使用语的用语集合。在手动索引(manual indexing)中，索引专家把能成为相应信息资料的提问语的用语作为索引词。

2.1 索引词的区别方法

资料中词语的出现频度用Zipf的法则把所有单词按出现频度顺序罗列时，出现频度与单词排位的倍数将是常数[4-6]。Luhn按出现频度区分有用的索引词和没用的索引词，去掉出现频度高的和出现频度低的单词。文件中出现的用语分为对文章有着重要意义的用语和不重要的、仅为构成文章需要的用语。英文文章中出现频度极高的冠词和前置词，虽然它们与文章的主要内容不相关，但是使用的频度极高。去掉出现频度极低的用语是因为其通常在对相应文章的主要内容没有相关性的假设下进行。

Luhn的索引词区别方法使用了英文文章的索引特征，即'所有的单词都会成为索引的候补'，'冠词、前置词、连接词和be动词等虽然出现的频度极高但对文章没有意义'。朝鲜语也可以把出现频度高但没有意义的助词、词尾、先语末语尾，及'-하다'，'-이다'等视为独立符号(token)的语句，以此作为朝鲜语的出现频度区分索引方法。但是朝鲜语和英语在索引词和非用词的区别方式中需要考虑本质上的区别。这要求在朝鲜语里区别索引词和非用词的方法中充分考虑朝鲜语的语言特征。

2.2 朝鲜语中的索引词和非用语

把高频度词语视为非用语的方法缩短了资料的占有量，从而有利于提高检索的速度。但是高频度词语根据意思模糊性或文脉也有索引价值，所以不能把所有高频度的词语视为非用语。如下30种词语在朝鲜语里的出现率特别高：

있다, 그, 수, 있는, 이, 것이다, 한, 한다, 및, 대한, 것이, 것은, 하는, 할, 그러나, 우리, 때, 등, 또, 같은, 것으로, 것을, 하고, 있었다, 그리고, 없는, 위한, 했다, 따라, 것

这些词大部分存在于所有文件中，其中'대 한'读法和写法都一样，但随文章的意思它本身的意思也会改变。

在英文里也有类似的词语，'vitamin A'的'A'在被视为非用语的情况下不会被检索，所以不能视为非用语。在英文中，这类单词都有可能成为'提问词'，比如最常使用的复合句'automatic indexing'，'named entity'等不仅名词、形容词，而且分词也会成为复合句的组合。在英文里名词、动词、形容词等一个单词的模糊词类的例子很普遍，大部分的词语都可以成为组成复合名词的要素，所以很难按照是否选择索引词或非用语进行分类。相反，朝鲜语的复合句都由名词组成，体词(名词、代词、数词等)和谓词因助词/词尾能简单地区分词类。所以，朝鲜语检索系统一般只把名词提取为索引词。

3 索引方法分析

对文件附加索引词的方法有词素索引、语句单位索引、n-gram索引和名词单位索引。词素单位索引是索引单词词素的方法，语句单位索引是将从文件中提取的语句直接作为索引词的方法，n-gram是以单词或者词素构成的部分句子作为索引单位的方法。索引词提取方法中有使用词素分析机和n-gram方法。为了补充n-gram法的缺点，先使用词素分析机，再对提取出来的索引词进行n-gram索引。

3.1 词素索引方法

词素单位索引法[17-21]是对代表文章的用语中提取词根(stem)作为索引词。这种方法通常包括删除非用语、词缀的切割和检出同一词根3个步骤。英文中也有因为去掉高频度用语导致索引词的个数减少30%～50%的情况。词缀切割是按照'同一的词根中引申的用语有着相同的内容，所以视为同一的词语'来切割扩展的词语。

英文中如nature, natures, natural, naturally, naturalness, naturalize等，都由一个词根扩展为不同的词类，对这样的词语一般用nature来索引。'相同词根检出'是英文中将类似stemming的单词变为'absorb'，'absorpt'为不同词素的情况下，为了索引成同一词根，把'-pt'结尾的字母改为'-b'，视为同一词根索引过程。

3.2 语句索引方法

语句索引[2,3]是为了补充词素索引法缺点而提出的方案。在相同的概念下把所扩展的词语索引成同一用语时，使用者检索的是'nature'，但会出现'naturally'的情况。像这样的词素单位索引不能满足使用者所需要的检索要求，所以在众多信息资料的Web检索系统中索引语句本身的方法最为合理。为了克服这样的非效率性，在语句索引法中不使用词素索引法的'词缀切割'和'检出同一词根'的过程，而是直接对语句本身进行索引。

Google对朝鲜语索引是用膠着性特性使用词素分析进行索引。把对应英语词缀的朝鲜语语法词素'助词/语尾/词缀'用词素分析法分解后把各个词素作为索引语，但会漏掉或不会复原非规则词素的语句。对变形语句进行索引时，按音节单位分解后提取索引词。Google的语句索引方式可达到原文一致的效果，但是如果提问词是'게'时，'어떻게/빠르게/짧게'的词语就不会被索引。

朝鲜语将题目作为检索词进行检索时，为了解决只用名词作为索引词时出现的问题，也使用了直接索引语本身的方法。此方法在一个语句中提取2个以上的索引词，词素分析的词类词素和语句本身都提取为索引词。但是多重语句索引法为了达到对少数提问的满足度和满足对所有的语句都要进行索引的要求，所以与英文文件语句索引有所不同，索引词的个数会增加，这是多重语句索引法的缺点。

3.3 N-gram索引方法

n-gram索引方法是在使用朝鲜语、汉字等2字节的语言系统中(double-byte code system)解决stemming与词缀切割等词素分析中提取的词根的误差问题，不经过词素分析便能对所有句子进行索引。主要方法如下：空格、特殊文字等作为区分字提取语句；在语句中将英文、数字和特殊文字等区分开来进行提取，在提取英语索引词时进行词根化(stemming)；在

提取的纯朝鲜语过程中把语句从第一音节连接一起的，按2个音节提取索引词。

N 为2的bigram方法最为常用，对连续2个字的用语进行全部索引'-하다'时，'정보'，'보검'，'검색'，'색체'，'체계'，'계를'会成为索引词。n-gram索引方法具有将所有句子进行检索的优点，但也会提取不需要的检索词，从而导致错误的索引结果。这种错误可通过只提取提问的开始或者结尾部分的方式来防止。但是，为了防止这类的次评委，需要有所有bigram索引中记录语句的开始-结束命令，为此索引的DB会增加。

n-gram索引法在索引时把提问句分为2个字，还对AND进行演算，所以会降低检索速度。因此，在'信息检索系统'中进行检索的bigram方式里至少需要5次对bigram句子的检索和对各检索结果的AND进行演算。n-gram索引法具有可以简化词素分解的优点，但具有如下缺点：索引DB过大；检索速度很慢；不能为使用者提供索引词。

改善n-gram索引方法有'使用词素分析后只对提取的索引词进行n-gram方法'和'词素分析中所提取出来的用于只对复合名词进行n-gram方法'。词类方法在词素分析的索引法中没有分解复合名词，而是用n-gram方式分解复合名词。

英语使用1字节(byte)文字的语言里隔写法很明确，不需要用这种方法，但朝鲜语和汉字中此方法会提高索引及检索的效率。n-gram方法初期在系统或研究开发中比较适用，但在商用系统中都采取了词素分析索引法。

3.4 名词单位索引法

名词单位索引[12,13]中的名词提取方法首先用分析排除信息的方法来排除词素分析，然后用语节提取名词，而对不能提取名词的语节可以通过词素分析过程提取名词。具体方法如下：利用分析排除信息的词素分析排除，对于不带名词的语节省略分析过程；利用后语节的名词抽取，检查搁在后语节前边的体词来提取名词；词素分析和声韵现象复原，通过后语节分析，当分析不出来时采用这种方法。例如'정보검색'这个语节能提取'정보'，'검색'，'정보검색'索引词。由此可见，名词提取不进行完全的词素分析，不如词素分析准确。

4 朝鲜语文件索引属性

4.1 朝鲜语的代表特征

(1)文章成分语调很顺畅。

(2)助词和语尾很发达，语型变化很大。

(3)体词和为此带上的助词和语尾决定文章成分，谓语一般放在句末。

(4)修饰词放在中心词的前边。

(5)体词和谓词的性质没什么区别。

(6)各词之间分写不是很明确。

4.2 索引语的抽取法

一般地，有两种索引提取方法：其一是众所周知的'stemming和后缀切断法'；其二是'词素的分离和变型法'。日语和汉语都不用分写，故单词分割(word segmentation)成了词素分析的核心。'stemming和后缀切断法'由两个过程构成：一是在朝鲜语中用空格来区分而去除对输入语节的非用语，二是切断stemming和后缀。而'词素的分离和变型法'中词素不需要用空格的区分连起来，所以识别词素分界是非常重要的因素。

上述两种方法对于语言分类体系中的屈折语(inflectional language)和膠着语(agglutinative language)有明显的区别。朝鲜语属于膠着语，故用'词素的分离和变型法'比较合适。但与那些不用分写的日语和汉语相比，从独立的语节中分离助词、语尾的过程可视为屈折语后缀切断，故'stemming和后缀切断法'更合适。

从两种索引语抽取的观点来看，朝鲜语属于边界语言。朝鲜语本身语节分解很明显，是以语根为中心来构成语节的，所以采用'stemming和后缀切断法'。但是对于复合名词分解和经常所犯的分写错误而言，采取'词素的分离和变型'法来抽取所标示的索引内容是比较理想的。

4.3 朝鲜语文件索引方式

朝鲜语文件索引方式区别于索引对象的索引是基准的设置方法，将什么类型的词汇视为非用语。朝鲜语的索引语和索引方式是以索引语为基准来选择的，有如下类型：语汇词素索引法，包括语汇词素索引和名词抽取索引；功能语索引法，包括词素原型索引和词素分离索引；合成词索引，包括合成词和构成要素索引、合成词组合索引和合成词分解索引。用各种索引法将下边的例句罗列出来。

例句：金策工业综合大学是位于风景优美的平壤大同江边的朝鲜尖端技术和信息的综合科学技术最高殿堂。

김책공업종합대학은 평양의 풍치수려한 대동강 기슭에 자리잡고 있는 첨단기술과 정보의 종합적인 과학기술의 최고전당이다.

(1)语汇词素索引结果

金策工业综合大学 平壤 风景 优美 大同江边 位于 尖端技术 信息 综合 科学技术 最高殿堂 是

김책공업종합대학 평양 풍치 수려하다 대동강 기슭 자리잡다 있다 첨단 기술 정보 종합적이다 과학기술 최고전당 이다

(2)名词索引结果

金策工业综合大学 平壤 风景 大同江 江边 尖端技术 信息 最高殿堂

김책공업종합대학 평양 풍치 대동강 기슭 첨단기술 정보 과학기술 최고전당

(3)词素原型索引结果

金策工业综合大学＋平壤＋风景优美＋大同江边＋位于＋尖端技术＋信息＋综合的＋最高殿堂＋是

김책공업종합대학+은 평양+의 풍치 수려하+ㄴ 대동강 기슭+에 자리잡+고 있+는 첨단기술+과 정보+의 종합적+인 과학기술 최고전당+이다。

(4)词素分离索引结果

上述4种索引抽取方式都不分解复合名词的构成要素。按照复合名词的索引方式，上述的例句复合名词'김책공업종합대학,첨단기술, 과학기술, 최고전당'索引抽取式如下。

合成词和构成要素的索引：

金策工业综合大学，金策，工业，综合，大学，金策工业，综合大学，工业大学，尖端，技术，尖端技术，科学，科学技术，最高，殿堂，最高殿堂。

김책공업종합대학,김책 , 공업 , 종합 , 대학 , 김책공업 , 종합대학 , 공업대학 , 첨단 , 기술 , 첨단기술 , 과학 , 과학기술 , 최고 , 전당 , 최고전당

合成词分解索引：

金策＋工业＋综合＋大学，尖端＋技术，科学＋技术，最高＋殿堂。

김책+공업+종합+대학 ,첨단+ 기술 , 과학+기술 , 최고+전당

'合成词和构成要素索引法'是索引出合成词本身和构成合成词的要素的索引法。所以这是一个把复合词和构成要素当成索引语取出的而对一个语节重复同一文字列的多重索引方法。与此相似的多重索引方法'复合名词组合索引法'是在复合词构成3个以上的要素时将组合各构成要素造出来的复合词作为索引词进行索引的。例如'정보검색시스템'：用'정보','검색','체계'组合'정보검색','검색 시스템'附加索引词。即使不是相邻的词也可以构造出'정보시스템'的词来索引。

'复合词组合索引法'采取把能检索出的复合词组合成预先全组合的方式，而'复合词分离索引法'是只索引出一些构成合成词的部分，故在处理过程中先将输入的文字列分解成词素。

5 一种朝鲜语信息索引方法

5.1 索引分析

(1) 复合名词分解。朝鲜语存在多种复合名词，所以与其只索引复合名词，不如检索被分解的名词。

(2) 外来语等推定名词的分解。外来语等推定名词和其他名词构成复合名词时会发生分不开推定名词的情况。外来语用朝鲜语写时会用到多个缀字，且意思一样的外来语等推定名词在发音扩张时会有好的效果。

(3) 名词分解后复合名词的添加或去除。复合词分解成名词以后去掉合成词时会有好的效果，添加时同样也会有好的效果。随着质问分解结果，当复合名词在3个以上时就添加合成词，否则去除。

5.2 信息索引

(1) 文献集合

使用了平壤信息开发中心发布的PIC 5.0里的提问集合与文献集合。提问词集合由整个提问中任意提取的50个单词和科学提问的30个单词组成，并且进行了索引评价。

(2) 检索模型

本文以按照索引语提取方法比较检索性能为目的，使用了一种加权值分配法。利用基于2-Poisson概率分布的TREC-8的Okapi系统使用的BM25方法进行加权值计算，如式(1)所示：

$$\mathrm{sim}(d_j,q)=\sum_{t_i\in d_j}\left(\frac{tf_i}{kl\cdot((1-b)+b\cdot\frac{dl_j}{avdl})+tf}\cdot\log\frac{N-df_i+0.5}{df_i+0.5}\right)\cdot qtf_i \quad (1)$$

式中，kl和b值使用了实验中结果最好的1.5，0.5数值(tf：索引词的频度数，dl：文献长度，$avdl$：文献的平均长度，df：文献频度，qtf：提问词中的索引词频度数，N：整个文献数)。

(3) 索引结果

4种索引法进行索引词提取的索引结果如表1所列。在分析过程中，对前10位文献的正确率(10_precision)和前1000位文献的召回率(Recall)进行评价。除了语句索引方法之外，其他方法的索引结果大概相似。利用语句索引法索引时，跟其他索引方法比较起来，不能索引单词中间的名词等等，因此索引的准确率下降。在名词索引法中，提问词越长，索引词提取的结果越好。而提问词长时，在n-gram索引方法中生成较多的非用词。在词素分析中，因复合名词单一索引词提取及分析方法本身的误差而生成较多的非用词。而提问词短时在题目中词素单位索引方法的结果最好。

表1 不同索引方法的索引结果

索引词	提问词	平均准确率	10_Precision	召回率
名词		0.2135	0.4580	0.5345
词素	TITLE	**0.2230**	**0.4980**	0.5296
n-gram	(题目)	0.2176	0.4560	**0.5502**
语节		0.1895	0.4300	0.5231
名词		**0.2598**	0.4920	0.6380
词素	DESC	0.2463	0.4920	0.6201
n-gram	(摘要)	0.2566	**0.5240**	**0.6387**
语节		0.2210	0.4780	0.5888

5.3 提出的索引方法

通过总结以上分析结果，本文提出如下朝鲜语信息索引方法，提取过程如图1所示。

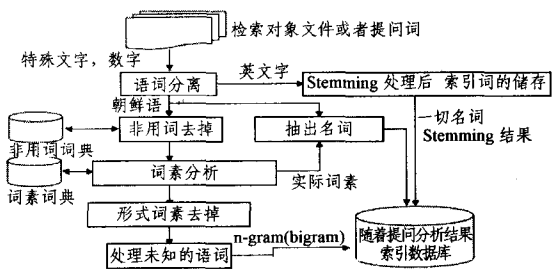

图1 提出的朝鲜语信息索引方法索引词提取过程

步骤1 空格，特殊文字等看成区分子抽取语节；

步骤2 通过区分词语中字母、数字、特殊文字等进行索引，并按提取的英文词处理Stemming；

步骤3 很难选为索引词的特殊文字和数字等都用非用语来去除；

步骤4 取出的语节用品词区分法对分解名词、收取名词等时机词素进行抽取；

步骤5 判断在步骤3中取出的实际词素是推定名词或动词索引；

步骤6 在步骤3中取出的实际词素有3个语节并且在不发生名次分解时对前两个语节进行索引；

步骤7 抽取的语节用名词抽取机抽选名词，当抽出的名词在步骤3中取出的实际词素里不存在长度为6字(12bite)以上的词时就附加为索引词。

以句子"n-grams 방법을 리용한 색인어 추출방법및 세계적인 추세에 대한 책을 요구 합니다."为例，分析如下：

(1) 语句分离

n-grams, 방법, 방법을, 리용, 리용한, 색인, 색인어, 추출, 추출방, 추출방법, 및, 세계, 세계적, 세계적인, 추세, 추세에, 대한, 책을, 요구, 합니, 합니다

(2) 英文字(Stemming)

按提取的英文字处理Stemming。n-gram(在词典上登

记)

(3) 非用词去掉

및, 대한 (去掉)

(4) 抽出名词

방법, 리용, 색인, 색인어, 추출, 추출방법, 세계, 추세, 책, 요구 (在词典上登记)

(5) 词素分析

방법+으+ㄹ, 리용하+ㄴ, 색인+어, 추출방법, 세계적+이+ㄴ, 추세+에, 책+으+ㄹ, 요구, 하+ㅂ니다

(6) 形式词素去掉

没有(如有未登记的词采用 n-gram 索引法处理)

(7) 采用提出的索引法提取索引词

n-gram, 방법, 리용, 색인, 색인어, 추출, 추출방법, 세계, 추세, 책, 요구, 세계적

(8) 名词单位索引法

n-grams, 방법, 리용, 색인, 색인어, 추출방법, 세계, 추세, 책, 요구

(9) 词素分析法

방법, 리용, 색인, 추출방법, 세계적, 추세, 책, 요구

(10) n-gram 索引法

n-grams, 방법, 법을, 리용, 용한, 색인, 인어, 추출, 출방, 세계, 계적, 적인, 추세, 세에, 대한, 책을, 요구, 합니, 니다

利用 4 种索引方法的索引词对提问及摘要进行实验,其结果如表 2 所列。从表 2 可以看出,提出的索引词提取新方法对题目及摘要的平均准确率最高。对于前 10 位文献的准确率,词素分析最好,并且对召回率而言,新方法最好。所提出的新方法具有通过语句分离及英文字复原、可提取准确的索引词、通过去掉非用词及形式词在检索中不利用非用词等优点。从上面分析可看出,提出的索引新方法与别的方法比起来,具有更高的准确率和召回率,因而具有更好的性能。

表 2 各种索引方法的索引结果

索引词	提问词	平均准确率	10_Precision	召回率
名词		0.2135	0.4580	0.5345
词素	TITLE	0.2230	**0.4980**	0.5296
n-gram	(题目)	0.2176	0.4560	0.5502
新方法		**0.2365**	0.4880	**0.5648**
名词		0.2598	0.4920	0.6380
词素	DESC	0.2463	0.4920	0.6201
n-gram	(摘要)	0.2566	0.5240	0.6387
新方法		0.2652	0.5160	0.6509

结束语 本文通过对名词单位索引、词素索引、n-gram 单位索引等典型索引方法进行比较、分析,深入剖析了影响检索性能的关键要素。同时,深入阐述朝鲜语的 30 个非用词、索引方式与朝鲜语的特征,进一步表明只有进行词素分析,才能在简单分析复杂的语词中提取出索引词。最后,通过结合各种索引方法的优点,提出了一种朝鲜语信息索引的新方法。仿真结果表明,所提出的索引新方法与别的方法比起来,具有更高的准确率和召回率,因而具有更好的性能。本文是根据复合名词的长度信息决定索引词附加的可能性,将来还会提出改进复合名词处理的新方法。

参 考 文 献

[1] Baeza-Yeates R, Ribeiro-Nero B. Modem Information Retrieval [M]. Addison-Wesley, 1999

[2] Park H-G, Ahn Y M, Seo Y-H. Korean Part-of-Speech Tagging System Using Resolution Rules for Individual Ambiguous Word [M]. 2007

[3] Choi W-J, Lee D-G, Rim H-C. Improving Korea Part-of-Speech Tagging Using The Lexicai Specific Classifier[M]. 2004

[4] Zipf H, Human P. Behaviour and the Principle of Least Effort [M]. Addison-Wesley, 1949

[5] Miller G A, Newman E B. Tests of a statistical explanation of the rank-frequency relation for words in written English[J]. American Jounal of Psychology, 1958, 71: 209-218

[6] Miller G A, Newman E B, Friedman E A. Length-frequency statistics for written English[J]. Information and Control, 1958, 1: 370-389

[7] Luhn. The Automatic Creation of Literature Abstracts[J]. IBM Jouurnal of Research and Development, 1958, 2: 159-165

[8] Cavnar W B. N-gram-based text filtering[C]// Harman D K, ed. Proceedings of the Second Text Retrieval Conference (TREC-2). National Institute of Standards and Technology, 1993: 171-179

[9] Lim H-S. An improved kNN learning based korean text classifier with heuristic information[J]. Digital Object Identifier, 2002, 2: 731-735

[10] Hong S-Y, Lee K-S, Rim S-K, et al. Customer satisfaction index measurement and importance-performance analysis for improvement of the mobile RFID services in Korea[C]// IEEE Conference. 2008: 2657-2665

[11] Lee B-H, Park D-W, Chung Y-J, et al. Efficiency improvement of Korean information retrieval system using relative distance between index words[C]// IEEE Conference. 2001: 243-246

[12] Kim J-H, Kwak B-K, Lee G, et al. A corpus-based learning method of compound noun indexing rules[J]. Korean Information Retrieval, 2001, 4(2): 115-132

[13] Li Qing, Chen Yuan-zhu, Myaeng S-H, et al. Concept unification of terms in different languages via Web Mining for information retrieval[J]. Information Processing and Management, 2009, 45(2): 246-262

[14] Kim M-S, Whang K-Y, Lee J-G. n-gram/2 L-approximation: a two-level n-gram inverted index structure for approximate string matching[J]. Computer Systems Science and Engineering, 2007, 22(6): 365-379

[15] Kim M-S, Whang K-Y, Lee J-G, et al. Structural optimization of a full-text n-gram index using relational normalization [J]. VLDB Journal, 2008, 17(6): 1485-1507

[16] Park J-H, Song Y-I, Rim H-C. Smoothing algorithm for n-gram model using agglutinative characteristic of Korean[C]// International Conference on Semantic Computing. 2007: 397-404

[17] Ali M N Y, Al-Mamun S M A, Das J K, et al. Morphological Analysis of Bangla Words for Universal Networking Language[C]// IEEE Conferences. 2008: 532-537

[18] Rahul C, Dinunath K, Ravindran R, et al. Rule-based Reordering and Morphological Processing for English-Malayalam Statistical Machine Translation[C]// IEEE Conferences. 2009: 458-460

[19] Hettige B, Karunananda A S. Web-based English-Sinhala translator in action[C]// IEEE Conferences. 2008: 80-85

[20] Durgar El-Kahlout I, Oflazer K. Exploiting Morphology and Lo-

cal Word Reordering in English to Turkish Phrase-based Statistical Machine Translation[C]//IEEE Early Access. 2009:1

[21] Hettige B, Karunananda A S. A Morphological Analyzer to Enable English to Sinhala Machine Translation[C]//IEEE Conferences. 2006:21-26

[22] Duwairi R, Al-Refai M, Khasawneh N. Stemming Versus Light Stemming as Feature Selection Techniques for Arabic Text Categorization[C]//IEEE Conferences. 2007:446-450

[23] Karageorgakis P, Potamianos A, Klasinas I. Towards incorporating language morphology into statistical machine translation systems[C]//IEEE Conferences. 2005:80-85

[24] Robertson S E, et al. Okapi at TREC-8, In The Enghth Text Retrieval Conference (TREC 8) [J]. Gaithersburg, MD: NIST, 2000

[25] 刘海峰,王元元,丘国防. 密度聚类模式下一种基于层次的自动文摘方法研究[J]. 情报杂志,2007(03)

[26] 刘金红,陆余良. 基于 Ontology 改进的 N-Gram 文本分类模型研究[J]. 计算机工程与设计,2007(13)

[27] 于津凯,王映雪,陈怀楚. 一种基于 N-Gram 改进文本特征提取算法[J]. 图书情报工作,2004,48(8):48-51

[28] 朱志国,邓贵仕,孔立平. 基于 N-gram 的 Web 用户浏览模式分类算法研究[J]. 情报学报,2009(6):389-394

[29] 刘鹏远,赵铁军. 基于 Web 的无指导译文消歧词模型与 N-gram 模型及对比研究[J]. 电子与信息学报,2009,31(12)

[30] 秦健. N-gram 技术在中文词法分析中的应用研究[D]. 青岛:中国海洋大学,2009:1-63

(上接第 131 页)

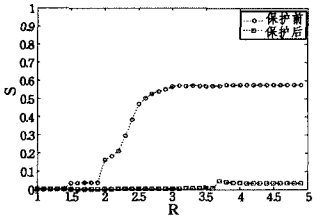

图 6 保护前后软件相继故障规模

由图 6 可知,在对关键节点提供保护之后,蓄意攻击所能造成的故障规模明显减小,控制在 5%左右。点强度的分布具有明显的无标度特性,也就是说点强度较大的节点只占节点总数的很少一部分。通过表 3 也可以看出,关键节点只定义了 10 个,而总的节点有 623 个,关键节点占总节点的 1.6%,这样就可以用较少的测试用例来保证软件的可信性。

结束语 本文把软件执行路径看作是加权网络来研究,利用复杂网络的理论来分析软件执行路径。在进行大量实验的基础上,发现软件执行路径的加权网络具有小世界效应和无标度特性,不仅仅点强度分布具有无标度特性,权重的分布同样具有无标度特性。

基于 CML 相继故障模型,研究软件系统中各个节点(函数)的非线性动力学特征、软件系统大规模故障的形成机理与传播行为。根据模拟结果可以看出,点强度越大,它发生故障所能造成的最终故障规模越大。因此本文提出了基于关键节点的软件测试技术,定义点强度大的节点为关键节点。基于 gedit 软件,模拟了基于关键节点的软件测试技术的效果。定义 1.6%的关键节点,就可以保证故障的最终规模在 5%以内,即使用较少的测试用例就能保证软件的可信性。

参 考 文 献

[1] 陈火旺,王戟,董威. 高可信软件工程技术[J]. 电子学报,2004,31(12):1934-1938

[2] 刘克,单志广,王戟,等. "可信软件基础研究"重大研究计划综述[J]. 中国科学基金-科学进展与展望,2008,3:145-151

[3] Watts J S, Strogatz S H. Collective dynamic of small-world networks [J]. Nature, 1998, 393(6684):440-442

[4] Strogatz S H. Exploring complex networks [J]. Nature, 2001, 410(6825):268-276

[5] Albert R, Barabási A-L. Statistical mechanics of complex networks [J]. Rev. Mod. Phys, 2002, 74(1):47-97

[6] Dorogovtsev S N, Mendes J F F. Evolution of Networks: From Biological Nets to the Internet and WWW [M]. London: Oxford University Press, 2003

[7] Newman M E J. The structure and function of complex networks [J]. SIAM Review, 2003, 45(2):167-256

[8] Potanin A, Noble J, Frean M, et al. Scale-free geometry in object-oriented programs [J]. Communications of the ACM, 2005, 48(5):99-103

[9] Valverde S, Ferrer-Cancho R, Sole R V. Scale-free Networks from Optimal Design [J]. Europhysics Letters, 2002, 60(4): 512-517

[10] Myers C R. Software systems as complex networks: structure, function, and evolvability of software collaboration graphs [J]. Phys. Rev. E, 2003, 68(4):046116

[11] Wheeldon R, Counsell S. Power Law Distributions in Class Relationships [C]//Proc Third IEEE Int'l Workshop Source Code Analysis and Manipulation. 2003

[12] Crucitti P, Latora V, Marchiori M. Model for cascading failures in complex networks[J]. Phys. Rev. E, 2004, 69:045104

[13] Kaneko K. Period-doubling of kink-antikink patterns, quasiperiodicity in antiferro-like structures and spatial intermittency in coupled map lattices [J]. Prog. Theor. Phys, 1984, 72(3): 480-486

[14] Kaneko K. Coupled map lattices[Z]. Singapore: World Scientific, 1992

[15] Concas G, Marchesi M, Pinna S, at al. Power-laws in a large object-oriented software system [J]. IEEE Transactions on Software Engineering, 2007, 33(10):687-707

[16] Hyland-Wood D, Carrington D, Kaplan S. Scale-free Nature of Java Software Package, Class and Method Collaboration Graphs [R]. MIND Laboratory, University of Maryland Collage Park, 2006

[17] Zheng X L, Zeng D, Li H Q, et al. Analyzing open-source software systems as complex network [J]. Physica A, 2008, 387(24):6190-6200

[18] 王宏霞,何晨. 细胞神经网络的动力学行为[J]. 物理学报,2003,52(10):2409-2414

[19] 陈星光,周晶,朱振涛. 基于耦合映像格子的城市交通系统相继故障研究[J]. 数学的实践与认识,2009,39(7):79-84