软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2011,22(8):1816−1826 [doi: 10.3724/SP.J.1001.2011.03890]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于概念和语义网络的近似网页检测算法*

曹玉娟[1,2+], 牛振东[1], 赵 堃[1], 彭学平[1]

[1](北京理工大学 计算机科学技术学院,北京 100081)

[2](北京航天飞行控制中心,北京 100094)

Near Duplicated Web Pages Detection Based on Concept and Semantic Network

CAO Yu-Juan[1,2+], NIU Zhen-Dong[1], ZHAO Kun[1], PENG Xue-Ping[1]

[1](School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

[2](Beijing Aerospace Command Centre, Beijing 100094, China)

+ Corresponding author: E-mail: cyjmdy@gmail.com

Cao YJ, Niu ZD, Zhao K Peng XP. Near duplicated Web pages detection based on concept and semantic network. *Journal of Software*, 2011,22(8):1816−1826. http://www.jos.org.cn/1000-9825/3890.htm

Abstract: Reprinting websites and blogs produces a great deal redundant WebPages. To improve search efficiency and user satisfaction, the near-Duplicate WebPages Detection based on Concept and Semantic network (DWDCS) is proposed. In the course of developing a near-duplicate detection system for a multi-billion pages repository, this paper makes two research contributions. First, the key concept is extracted, instead of the keyphrase, to build Small Word Network (SWN). This not only reduces the complexity of the semantic network, but also resolves the "expression difference" problem. Second, this paper considers both syntactic and semantic information to present and compute the documents' similarities. In a large-scale test, experimental results demonstrate that this approach outperforms that of both I-Match and keyphrase extraction algorithms based on SWN. Many advantages such as linear time and space complexity, without using a corpus, make the algorithm valuable in actual practice.

Key words: duplicate removal algorithm; small world network; near duplicated Web page; standard deviation

摘 要: 在搜索引擎的检索结果页面中,用户经常会得到内容近似的网页.为了提高检索整体性能和用户满意度,提出了一种基于概念和语义网络的近似网页检测算法 DWDCS(near-duplicate webpages detection based on concept and semantic network).改进了经典基于小世界理论提取文档关键词的算法.首先对文档概念进行抽取和归并,不但解决了"表达差异"问题,而且有效降低了语义网络的复杂度;从网络结构的几何特征对其进行分析,同时利用网页的语法和结构信息构建特征向量进行文档相似度的计算,由于无须使用语料库,使得算法天生具有领域无关的优点.实验结果表明,与经典的网页去重算法(I-Match)和单纯依赖词汇共现小世界模型的算法相比,DWDCS 具有很好的抵抗噪声的能力,在大规模实验中获得了准确率>90%和召回率>85%的良好测试结果.良好的时空间复杂度及算法性能不依赖于语料库的优点,使其在大规模网页去重实际应用中获得了良好的效果.

关键词: 网页去重算法;小世界网络;近似网页;均方差

中图法分类号: TP391 **文献标识码**: A

* 基金项目: 国家自然科学基金(60803050, 60705022); 新世纪优秀人才计划(NCET-06-0161)
收稿时间: 2009-10-09; 修改时间: 2010-01-20; 定稿时间: 2010-04-27

据中国互联网络信息中心(China Internet Network Information Center,简称 CNNIC)[1]2005 年公布的互联网统计报告表明,用户使用搜索引擎遇到的最大问题是重复信息过多.据统计,目前 Internet 上内容近似的网页约占 30%~45%[2],它们大多来自镜像网站或网站间的转载.这些冗余的信息不但浪费了大量存储资源,降低了索引效率,直接影响搜索引擎的整体性能,而且加重了用户的阅读负担.面对海量的数据,用户不愿意看到一堆内容重复和近似的信息;另一方面,被频繁转载的网页往往比较重要和热门,近似网页的发现也有助于为网页的排序提供依据.本文将以去除内容近似网页为目标,提出一种基于概念和小世界网络的特征提取及近似网页检测算法.

我们常把句法、结构完全相同的文档视为重复文档,但这里所指的近似网页是指正文内容基本相同的网页,而无论其句法、结构是否完全一致.

对于重复文本的检测采用传统的剽窃检测技术很容易实现,但由于网络噪声(广告、超链、编辑信息等)的影响,对于内容近似文档的检测就不那么容易了.本文提出的基于概念和小世界网络的近似网页去重算法 DWDCS(near-duplicate webpages detection based on concept and semantic network)兼顾考虑网页的语法和语义信息,对文档的概念进行提取;基于文本的小世界特性,构建文本概念共现网络,从网络的几何特征进行分析,抽取关键概念;与文档结构信息一起构建特征向量进行近似网页的去除.在进行特征提取时,通过概念归并和利用文本概念共现网络中节点度分布的幂率特性,有效提高了系统的运行效率.

本文第 1 节介绍经典的文本去重算法.第 2 节介绍小世界模型.第 3 节介绍我们对经典基于小世界模型的关键词提取算法的改进及论述 DWDCS 算法设计.第 4 节是实验结果和对结果的分析.第 5 节是结论.

1 现有近似网页检测算法介绍

近年来,针对近似网页的检测展开了许多研究,如网页结构近似性检测[3]、超级链接近似性检测[4]等.但本文关注的是内容近似网页的检测.

鲍军鹏等人[5]对自然语言文档复制检测研究进行了综述,包括基于字符串匹配的方法来度量文件之间的相似性的 sif[6]工具、基于向量空间模型和词频统计方法度量文本相似性的 SCAM(stanford copy analysis method)[7]及利用后缀树(suffix tree)来搜寻字符串之间的最大子串的 MDR(match detect reveal)[8].

我们常把文本复制检测算法分为两类:基于语法的方法(基于 Shingle 的算法)和语义的方法(基于 Terms 的算法).其中,Shingle 是指文档中若干个连续出现的单词.这种方法从文档中选取一系列 Shingle 后统计相同的 Shingle 数目或者比率,作为判断文本相似度的依据.文献[9−12]都是常用的基于 Shingle 的算法.对于各种基于 Shingle 的算法,Ye[13]就其参数选择进行了实验分析;基于 Terms 的方法采用单个词条作为计算的基本单元,而不考虑词条出现的位置和顺序.其中最著名的就是 Chowdhury 的 I-Match[14]算法,选取 IDF 值(inverse document frequency)较高的词条排序后构成为文档的指纹(fingerprint),指纹相同的文档被视为内容近似网页.其他基于 Terms 的方法[15−17]也大都采用 SVM 模型,利用 TF/IDF 值进行文档关键词的提取,并将关键词作为文档的特征向量,通过计算文档间的相似度来进行近似网页的检测.

基于 Shingle 的方法基于精确匹配,适用于重复文档的排查,而对于含有噪声的内容近似网页,往往会漏检;而基于 Terms 的方法依赖于语料库,缺乏深层挖掘文档本身的语义特征,关键词自动标引的准确性和有效性都有待提高,而且由于其依赖语料库,导致了其领域相关性的缺点.

基于文档的语义网络(semantic network)可以更好地对关键词和作者想要表述的重要概念进行提取.我们将采用基于概念和语义网络的方法进行文档关键概念的提取,结合文档结构信息构造特征向量,用于内容近似网页的检测.

2 小世界模型

2.1 小世界网络

小世界现象源于社会学家 Milgram 于 1967 年开展的有关追踪美国社会网络中最短路径的研究.研究结果

表明,任意一对美国人之间,大都可以找到不多于 6 个两两相识的人将他们联系起来,这就是著名的“六度分离”(six degree separation)问题.小世界现象目前还没有精确的定义,通常认为,如果网络中两节点间的平均距离 L 随网络节点数目 N 成对数增长,即 $L\propto \ln N$,则称该网络具有小世界现象.

Watts 于 1998 年在 Nature 杂志上发表的论文对小世界现象进行了深入研究[18],提出小世界网络具有高聚度和短路径的特性.

2.1.1 聚 度

聚度(cluster coefficient)是网络中两点间两两相连的比例,反映了网络结构局部特征.对于节点 i,假设其有 k_i 个邻居节点(k_i 也称为节点 i 的度),则这 k_i 个节点最多存在 $k_i\times(k_i-1)/2$ 条边(这种情况仅仅发生在所有节点两两相连的时候).φ_i 为 k_i 个邻居节点间存在的实际边数.节点 i 的聚度为 C_i:

$$C_i=\frac{\varphi_i}{k_i\times(k_i-1)/2} \tag{1}$$

整个网络的聚度为 C:

$$C=\frac{1}{N}\sum_{i=1}^{N}C_i \tag{2}$$

具有小世界性质的网络具有聚度高的特性.

2.1.2 平均路径长度

路径长度(path length)是指网络中任意两点间最短距离的平均数,它表现了网络结构的全局特征.对于节点 i,其平均路径 d_i 为

$$d_i=\frac{1}{L-1}\left(\sum_{j=1,j\neq i}^{L}d_{\min}(i,j)\right) \tag{3}$$

整个网络的平均路径长度 d 为

$$d=\frac{1}{L}\left(\sum_{i=1}^{L}d_i\right) \tag{4}$$

具有小世界性质的网络的平均最短路径会很短,远小于网络规模[18](这也是“小世界”命名的原因).

2.1.3 节点度的幂率分布特性

Yutaka[19]和 Ferrer[20]等人指出,人类的语言及由文档构成的词汇共现网络同样具有高聚度、短路径的特性.Ferrer[20]的研究还表明:词汇共现网络还具有无标度(scale-free)特性;网络中节点度的分布接近于幂率分布(power law),即,度为 k 的节点在网络中的概率 $Pr(k)\propto k^{-r}$,r 为常数.它反映了网络中各节点之间的连接状况(度数)具有严重的不均匀性,只有很少数的节点与其他节点有很多的连接,成为“中心节点(hub node)”,而大多数节点的度很小,即存在所谓的“长尾(long tail)现象”.中心节点对无标度网络的运行起着主导作用.

近年来,运用小世界理论对各种复杂网络(运输网络、疾病传播、互联网控制等)的研究不断涌现.

2.2 经典基于小世界理论的文档关键词提取算法

Yutaka[19],Zhu[21]和 Huang[22]等人利用小世界特性构建了语义网络,进行文档关键词的提取.基于小世界特性进行文档关键词的提取的方法,是从网络结构的几何特征来分析文本词汇的语义网络;利用深层的语义信息,通过计算词汇对语义网络的贡献程度,获取文档的主题关键词,具有不依赖于语料库的优点.

但大部分基于小世界理论提取关键词的算法是通过词汇构建语义网络,没有进行概念的提取,因此网络结构非常复杂;加上没有利用网络节点度分布的无标度特性,导致算法的时间复杂度太高($O(n^3)$)而无法大规模用于工程.虽然 Huang 采用邻接表存储结构和缓存迭代中间结果的方法在一定程度上降低了时间和空间复杂度,但它更适用于对时间效率要求比较宽松的数字图书馆系统.对于网页数据,虽然重复网页的排查可以离线进行,但海量的数据量对算法的时空特性提出了更高的要求,经典的基于小世界理论的方法已无法满足需求.

另外,现有的基于小世界理论提取关键词算法大都没有考虑表达差异问题,即在人类的自然语言中,随着时间、地域或领域的改变,甚至在同一篇文章中,作者为了表达丰富,也会用不同的语言表达形式来表达同一概

念(如,后面的实验用例“Foods that could trigger a nasty headache”中的citrus,orange和tangerine;deficiency和lack等).表达差异问题会带来语义网络Hub节点的分散,影响文档关键词提取的精度.

我们对经典基于小世界理论提取文档关键词的算法主要做如下改进:(1) 对文档概念进行抽取和归并,有效解决了表达差异问题.而且,概念的归并相当于首先对网络中的节点进行聚类,把同一概念的节点合为一个,能够明显降低语义网络的复杂度.(2) 利用节点度的幂率分布特性和“中心节点对无标度网络的运行起主导作用”的研究结论,仅将度数高的节点作为候选节点(而不是对所有的节点),计算其对小世界网络的贡献率,以识别文档的关键概念,从而大量缩短了系统的运算时间.

3 基于概念和语义网络的去重算法设计

我们首先进行文档概念的提取,构建文档的概念共现语义网络;依据小世界现象中的幂率分布(power law)特性,计算Hub节点对小世界网络的贡献率,提取文档关键概念;然后,结合其位置和权重信息,构建特征向量进行文档表示;最后计算文档相似度,进行内容重复网页的检测.具体设计方案如下:

3.1 网页文本提取及预处理

网页中包含的广告、链接等噪声信息会对该网页内容检索产生干扰.因此,在对网页的内容建立索引之前,我们需要对其中的有效正文信息进行提取.这里采用的是我们另一项课题的研究成果[23],预处理包括短语识别[24]、停用词去除和词根还原等.

3.2 文档关键概念提取

3.2.1 基于Wordnet的概念归并

概念是客观事物的特有属性(或称本质属性)在人们头脑中的反映.把所感知的事物的共同本质特点抽象出来加以概括,就成为概念.用概念来表示文本,不但可以准确地表示文本的本质内容,而且利用概念的抽象性将数个同义词归结为一个概念,可以有效地降低小世界网络中节点的数目.WordNet使用同义词集合(synset)代表概念(concept).对于一个单词,其概念可以表示为它在WordNet中同义词的集合.出现词t的同义词时就可以用其概念$Con(t)$来代替同义词和词t本身.

$$Con(t)=\{st_i\}\ st_i \in synonymies\ of\ t \tag{5}$$

对于由多个单词$\{t_1,t_2,t_3,\ldots,t_n\}$组成的短语$p_i$,其概念$Con(p_i)$为组成短语单词的概念集合.

$$Con(p_i)=\{st1_i,st2_i,..,stn_i,\}\ stk_i \in synonymies\ of\ t_k \tag{6}$$

我们用短语间同义词重叠的比例来计算短语的概念相似度.

$$Rel(Con(p_1),Con(p_2))=\frac{number\ of\ overlap\ tokens\ in\ con(p_1)\ and\ con(p_2)}{(number\ of\ tokens\ in\ Con(p_1)+number\ of\ tokens\ in\ Con(p_2))/2)} \tag{7}$$

如果$Rel(Con(p_1),Con(p_2))>Rel_{thr}$($Rel_{thr}$为用户设定的阈值),则认为两个短语表达的概念相同,将其合并.对实验数据集(1 034 403篇网页)的统计表明,进行概念归并后,平均每篇网页概念的数量约为短语数量的63%.概念的归并不仅解决了表达差异问题,而且有效降低了小世界网络的复杂度.

3.2.2 构建概念共现图

概念提取后,采用如下步骤构建概念共现图:

(1) 节点的选取.选取在文档中出现频率$f>f_{thr}$的概念Con_i(f_{thr}为事先给定阈值)作为概念共现图的节点.

(2) 边的选取.Lyon[25]的研究表明,70%的语法关系存在于相邻短语间,17%的语法关系存在于距离不超过2的短语间.从降低算法复杂性和保留大部分短语间语法关系的角度考虑,我们为距离不超过2的概念(节点)建立边的联系.

(3) SWN(small word network)网络的定义.文档的概念共现网络即可定义为$GL=(N_L,E_M)$,其中,$N_L=\{Con_i\}$为节点(概念)的集合,$E_M=\{\{Con_i,Con_j\}\}$为边(概念关系)的集合.$\xi_{i,j}=\{Con_i,Con_j\}$表示节点$Con_i$与$Con_j$之间是否存在边,如果存在,则$\xi_{i,j}=1$;否则,$\xi_{i,j}=0$.

(4) 网络的裁剪.经统计发现,对于每篇文档,按照以上步骤构建的图,平均 96%的节点是相互连通的,构成一个大的连通主部;但仍有 4%的节点与其是非连通的.对此,Strogatz[26]也指出,这是复杂网络的一个普遍现象.Huang[22]的实验结果表明,直接裁剪掉这些节点不会影响网络的小世界特性.这些非连通节点在网页中多为转载信息、版权信息等,与文档主题没有直接关系.为此,我们直接剪裁与连通主部分离的那些节点.

对去重实验数据中的某篇网页“Foods that could trigger a nasty headache”(http://shine.yahoo.com/channel/health/foods-that-could-trigger-a-nasty-headache-507285/)构建其概念语义共现网络,如图 1 所示.

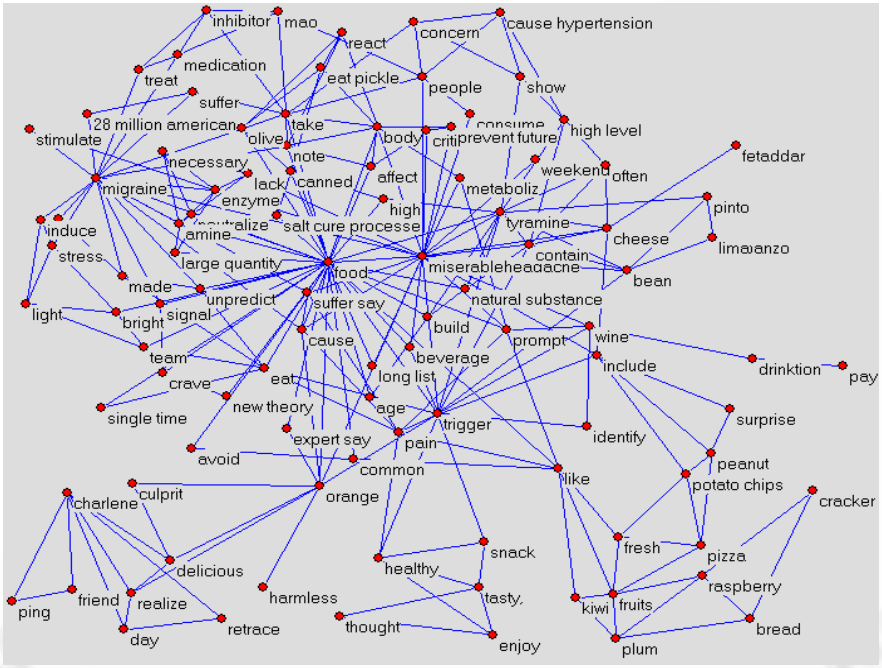

Fig.1 Concepts co-occurance network of webpage ‘Foods that could trigger a nasty headache’

图 1 网页“Foods that could trigger a nasty headache”概念共现网络

3.3 关键节点识别

经过分词、停用词去除等预处理后构建概念语义共网络 CN,对于 CN 可计算它的两个基本特性:聚度 C 和特征路径长度 d.

对于度为 k 的节点(概念)$Con_i \in TL$,其邻居节点为 $\Gamma_i=\{j|\xi_{i,j}=1\}$,邻居节点间存在的实际边数为

$$\varphi_i = \sum_{m,n=1}^{k} \xi_{m,n} \mid Con_m, Con_n \in \Gamma_i.$$

用公式(3)计算 CN 的特征路径长度.

设 CN_i 为移除第 i^{th} 个节点后的概念共现网络,$d_{iabsent}$ 为 CN_i 的平均路径长度.节点 i^{th} 对网络 CN 呈现小世界特征的贡献率为 $CB_i=d_{iabsent}-d$.CB_i 的值越大,表示该概念对文章结构的连接起关键作用.它们通过建立“捷径”为文档中的其他概念建立桥梁,一旦失去它们,文章就会按照不同主题细分成为一个个毫无联系的小网络,结构将会变得松散.

虽然概念的归并能够有效降低网络的复杂度,但是计算网络的最短路径仍然是一个十分耗时的工作.面对节点较多的长网页,该算法将由于时间复杂度太高而无法得到实际应用.我们利用概念共现网络中节点度分布的无标度(scale-free)特性和“中心节点对无标度网络的运行起主导作用”的研究结论[27],选取度 k 最大的 l 个节点(实验中 l=30,实验 4 将对 l 的取值进行论证)计算其 CB_i,按大小排序,抽取排在前面的 p 个概念作为关键概念

(在实际应用中,我们选取 p=20,实验 3 将对关键概念的数目进行论证).

3.4 特征向量构建

但是,仅仅使用关键概念是不够的,Hoad 和 Zobel[28]指出,传统的 TF/IDF 值进行余弦相似度计算,不足以用于文档相似度的检测.

关键概念在文档中的位置对于近似文档的检测也很重要.因此,我们利用向量列表 $Vp=(Lp_1,\ldots,Lp_i\ldots,Lp_{20})$, $Lp_i=(Posi_{i,l},\ldots,Pos_{i,j},\ldots,Pos_{i,n})$来记录特征项的位置.其中,$Pos_{i,j}$ 是第 i 个概念在文章中第 j 次出现的位置.

$$V_p=\begin{bmatrix} LP_1 \\ \ldots \\ LP_N \end{bmatrix}=\begin{bmatrix} Pos_{1,1} & \ldots & Pos_{1,m} \\ \ldots & \ldots & \ldots \\ Pos_{N,1} & & Pos_{N,k} \end{bmatrix} \tag{8}$$

对去重实验数据中的网页“Foods that could trigger a nasty headache”提取特征向量,得到最能表示该网页的 20 个关键概念为 food,miserable,headache,migraine,trigger,tyramine,cheese,orange,eat,body,take,pain,contain, enzyme,bean,fruit,cause,wine,like,health.

这意味着我们采用 20 个关键概念和它们的位置一起构成特征向量来表示一篇文档.

3.5 特征向量的存储和检索

为了对特征向量进行快速访问,可以将特征向量映射到 HASH 表中,但 HASH 表的查找比较适合于精确匹配.由于网页噪声的影响,即使是重复网页的文本特征向量有时也不完全相同,精确匹配会导致排查的失败.考虑到倒排索引具有实现相对简单、查询速度快、支持布尔查询等优点,我们使用倒排索引为特征向量进行存储和检索.

特征向量是最能代表一篇文章的一组概念及其位置特征,只须检索排在前边的 n 维特征向量并计算其相似度,即可基本确定两篇文章是否是近似文档.在得出匹配检索后,采用余弦公式(9)进行相似度计算:

$$\xi=\frac{d1\times d2}{\|d1\times d2\|}=\frac{\sum_{i=1}^{m}d1(i)\times d2(i)}{\sqrt{\sum_{i=1}^{m}d1(i)^2}\times\sqrt{\sum_{i=1}^{m}d2(i)^2}} \tag{9}$$

若$\xi>Sim_{thr}$(Sim_{thr}为用户自定阈值,实验中我们选取 Sim_{thr}=0.9),则可以推断 $d1$,$d2$ 很有可能是近似网页.两个网页特征项的距离矩阵可以表示为

$$V_{p_1}-V_{p_2}=\begin{bmatrix} Pos^1_{1,1}-Pos^2_{1,1} & \ldots & Pos^1_{1,1}-Pos^1_{1,m} \\ \ldots & \ldots & \ldots \\ Pos^1_{N,1}-Pos^2_{N,1} & \ldots & Pos^1_{N,1}-Pos^2_{N,m} \end{bmatrix}=\begin{bmatrix} \delta P_{1,1} & \ldots & \delta P_{1,m} \\ \ldots & \ldots & \ldots \\ \delta P_{N,1} & \ldots & \delta P_{N,m} \end{bmatrix} \tag{10}$$

其中, $Pos^1_{i,j}$ 表示第 1 篇文章中第 i 个关键概念第 j 次出现的位置; $Pos^2_{i,j}$ 表示第 2 篇文章中第 i 个关键概念第 j 次出现的位置;δP_{ij} 为两篇文档第 i 个关键概念第 j 次出现的位置之差,即距离.

第 i 个关键概念的平均距离 AVG_i 为

$$AVG_i=\frac{\sum_{j=1}^{r}|Pos^1_{i,j}-Pos^2_{i,j}|}{r}=\frac{\sum_{j=1}^{r}|\delta P_{i,j}|}{r} \tag{11}$$

通过方差 S_i 计算第 i 个关键概念距离的分布为

$$S_i=\frac{\sqrt{\sum_{j-1}^{r}(\delta P_{i,j}-AVG_i)^2}}{r} \tag{12}$$

对于整篇文档,N 个关键概念的距离分布为

$$S=\frac{\sum_{i=1}^{N}S_i}{N} \tag{13}$$

如果 $S<S_{thr}$,则说明关键概念在两个网页中出现的位置和顺序相似,此时可以确定两个网页为近似网页.综上所述,判断网页 A 是否与已建索引库中的网页重复,步骤如下:

1. 提取网页 A 的文本 Ta.
2. 对文本进行概念抽取,构建语义网络.对文档关键概念进行抽取,与其位置信息一起构成文本特征向量 Va.
3. 选取 Va 的前 n 维为查询输入条件,检索的特征向量索引库 $IDXVa$:
4. IF 有 k 个匹配输出 $VD_i(i=1,2,\ldots,k,k>0)$
5. For (i=1; i<k; i++) {
6. 计算 Va 与 VD_i 的相似度ξ
7. IF (ξ>阈值) {
8. 计算特征向量的距离分布 s
9. IF (s<阈值) {
10. A and D_i 是近似网页
11. Break;
12. }
13. }
14. }
15. ELSE
16. 未找到重复网页
17. 将 Va 增量建入特征向量索引库 $IDXVa$ 中.

4 实验结果及对比分析

为了评价本算法的正确性和效率,本文设计了一系列实验,其中,实验 1、实验 2 是对算法的正确性和时空效率进行验证,实验 3、实验 4 是对算法中参数关键概念个数 p 和的选择进行实验.

正确性是算法的生命,这里给出两个评价标准:重复网页召回率(recall)和去重准确率(precision),定义如下:

$$Recall=\frac{\text{正确去重的网页数}}{\text{存在的重复网页数}},\ Precision=\frac{\text{正确去重的网页数}}{\text{所有去除的网页数}}.$$

实验 1. 算法的正确性

为了检测 DWDSC 的性能,我们在军事、医学和计算机这 3 个领域选择了 72 个查询词,用 Google 检索查询词.在每组检索结果中选取内容相同或相似的网页共计 5 835 篇,并将这些近似网页插入已存在的文档集(包含 1 028 568 个网页)中.分别运行 I-Match(同样选取 20 个特征词)、单纯依赖词汇共现小世界模型的网页去重算法(DWSWN)和 DWDCS 算法进行近似网页检测.

在军事领域输入 23 个查询,实验结果如图 2、图 3 所示.

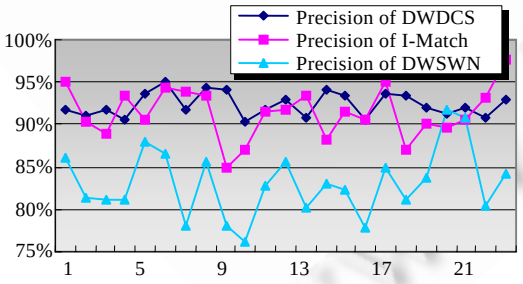

Fig.2 Precision in military field

图 2 军事领域准确率

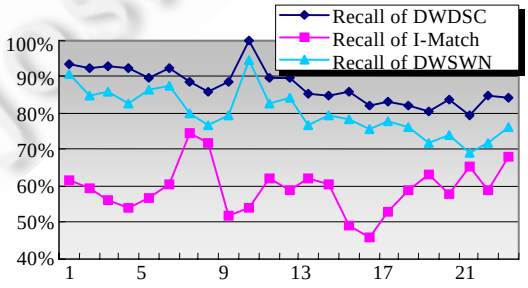

Fig.3 Recall in military field

图 3 军事领域召回率

医学领域输入 28 个查询,其中,20 组对应知识介绍性网页,8 组对应新闻性网页.实验结果如图 4、图 5 所示.计算机领域输入 21 个查询,全部对应新闻性网页,实验结果如图 6、图 7 所示.

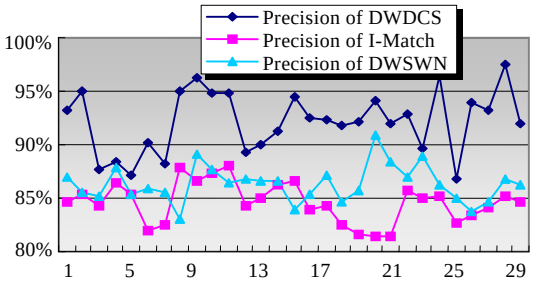

Fig.4 Precision in medical field

图 4 医学领域准确率

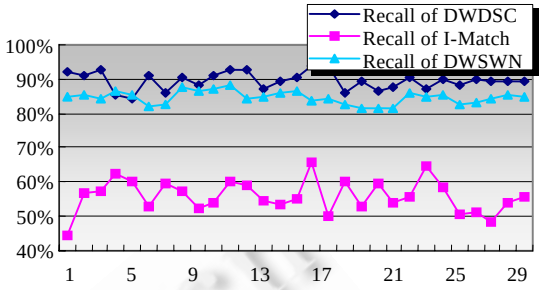

Fig.5 Recall in medical field

图 5 医学领域召回率

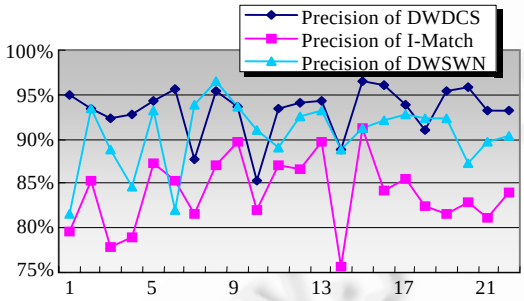

Fig.6 Precision in computer science field

图 6 计算机领域准确率

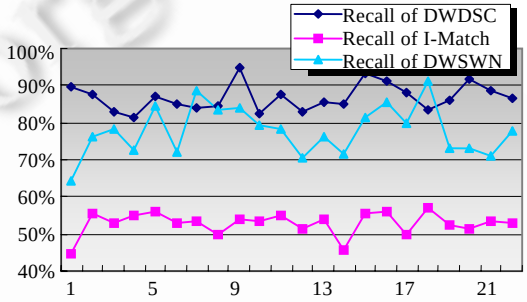

Fig.7 Recall in computer science field

图 7 计算机领域召回率

以上实验表明,DWDCS 算法在各个领域的文本去重实验中都表现出较高的准确率(>90%)和召回率(>85%),I-Macth 算法仅在军事领域表现出与 DWDCS 相当的准确率.其原因是,I-Macth 算法特征向量的抽取依赖于语料库中的词频,而本文实验用的是军事领域语料库,在该领域,I-Macth 表现出较好的抗噪能力;但在其他两个领域,准确率和召回率都有明显下降.这说明,I-Macth 算法具有领域相关的弱点.DWDSC 算法基于概念语义网络和文档的结构进行特征向量提取,与训练语料无关,算法具有领域无关的优点,因此在实验选取的 3 个领域都获得了更好的准确率和召回率.虽然 DWSWN 也具有领域无关的优点,但由于表达差异问题,准确率和召回率都逊于 DWDSC.同时我们也发现,3 种算法均在知识性介绍网页上取得了比新闻类网页更高的召回率,这是由于知识性介绍性网页通常包含噪音信息较少,特征向量的提取更为准确的缘故.

实验 2. 时空性能比较

用本文的方法(DWDSC)和 Yutaka[19]的方法进行语义网络构建,表 1 为网络节点数目及关键概念/关键词提取时间比较实验数据.

Table 1 DWDSC vs. Yutaka[17] in node number of the semantic network and average time for keyphrase extraction

表 1 DWDSC 与 Yutaka[17]的方法构建语义网络的节点数及关键节点提取时间比较

Characters in document	Number of documents	Average node number in the network		Average time for keyphrase extraction (ms/doc)	
		DWDSC	Yutaka	DWDSC	Yutaka
<500	163 361	145	227	1 685	2 379
500~2 000	814 336	498	786	6 931	32 579
2 000~5 000	54 395	1 219	1 938	24 377	184 342
>5 000	2 311	3 197	5 378	256 328	2 798 874

表 1 显示出,DWDSC 构建的基于概念的语义网络比起经典的词汇共现网络,节点数目平均降低了约 37%;利用无标度特性进行关键节点提取,时间平均减少了约 77%.

实验 3. 选取关键词的数量的讨论

如图 8 所示,实验结果表明,选取 20 个关键词可以获得更高的准确率和召回率.

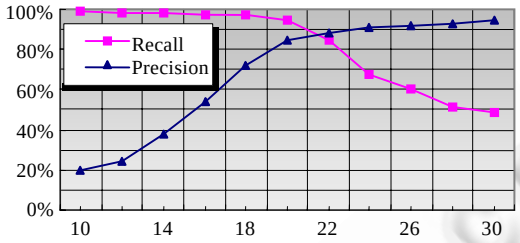

Fig.8 Selection on the number of the key concept

图 8 关键概念数量的选取

实验 4. l 取值的讨论

在第 3.3 节中我们提到,利用概念共现网络中节点度分布的无标度(scale-free)特性和“中心节点对无标度网络的运行起主导作用”的研究结论,选取度 k 最大的 l 个节点,计算其对小世界网络的贡献.本实验对 l 的取值进行讨论.在重复网页中随机抽取 100 篇,对 l 的不同取值分别计算关键概念提取时间、去重的准确率和召回率,得到如图 9 所示的曲线.

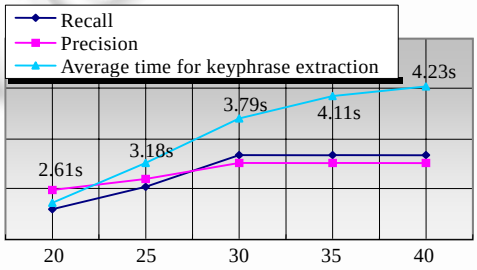

Fig.9 Selection on the parameter l

图 9 参数 l 的选择

图 9 显示,关键概念提取的平均时间随 l 呈近似线性的增长;但当 l≥30 后,网页去重的准确率和召回率无明显上升.因此,在实际应用中我们选取 l=30.

5 结论及未来工作

影响网页去重准确性的主要因素是网页噪声.本文提出的基于概念和语义网络的近似网页检测算法,能够有效减少噪声信息对算法准确性的不良影响;不依赖于语料库,基于概念和语义网络,提取文档关键概念;不但考虑了网页文本的内容、结构信息,而且充分利用检索系统的优势,获得了去重准确率>90%、平均召回率>85%的良好效果.本文的研究成果已成功应用于“国防科技工业网络信息资源信息采集系统”和“中国教育电视台学习超市平台建设项目统一检索系统”两个大型项目.

概念归并虽然有效降低了语义网络的复杂度,使得关键概念的提取速度得到显著提高,但计算节点对语义网络的贡献仍然是一项很耗时的工作.另外,对于短小网页,各概念在语义网络中的特征相似性导致关键概念提取的困难.关键概念在检索系统、文本聚类和分类系统中发挥重要作用,今后的工作包括语义网络关键节点识别时间性能的优化(尤其对于长文档)和短小文档关键概念的提取.

References:

[1] China Internet Network Information Center. The 16th statistics report on China Internet development. 2007 (in Chinese). http://www.cnnic.net.cn/index/0E/00/11/index.htm

[2] Cho JH, Shivakumar N, Garcia-Molina H. Finding replicated Web collections. Proc. of the ACM Int'l Conf. on Management of the Data, 2000,29(2):355−366. [doi: 10.1145/342009.335429]

[3] Li Z, Ng WK, Sun AX. Web data extraction based on structural similarity. Knowledge and Information Systems, 2005,8(4): 438−461. [doi: 10.1007/s10115-004-0188-z]

[4] Dean J, Henzinger MR. Finding related pages in the World Wide Web. In: Proc. of the 8th Int'l World Wide Web Conf. (WWW). Toronto: Elsevier, 1999. 1467−1479. [doi: 10.1016/S1389-1286(99)00022-5]

[5] Bao JP, Shen JY, Liu XD, Song QB. A survey on natural language text copy detection. Journal of Software, 2003,14(10): 1753−1760 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/14/1753.htm

[6] Manber U. Finding similar files in a large file system. In: Proc. of the 1994 Winter USENIX Technical Conf. San Francisco, 1994. http://manber.com/publications.html

[7] Shivakumar N, Garcia-Molina H. SCAM: A copy detection mechanism for digital documents. In: Proc. of the 2nd Int'l Conf. in Theory and Practice of Digital Libraries (DL'95). Austin, 1995. http://infolab.stanford.edu/~shiva/publns.html

[8] Monostori K, Zaslavsky A, Schmidt H. Document overlap detection system for distributed digital libraries. In: Proc. of the ACM Digital Libraries (DL2000). San Antonio, 2000. 226−227. [doi: 10.1145/336597.336667]

[9] Heintze N. Scalable document fingerprinting. In: Proc. of the 2nd USENIX Electronic Commerce Workshop. Oakland, 1996. 191−200.

[10] Krishna B, Broder A, Dean J, Henzinger MR. A comparison of techniques to find mirrored hosts on the WWW. Journal of the American Society for Information Science, 2000,51(12):1114−1122. [doi: 10.1002/1097-4571(2000)9999:9999<::AID-ASI1025>3.0.CO;2-0]

[11] Wu PB, Chen QX, Ma L. The study on large scale duplicated Web pages of Chinese fast deletion algorithm based on string of feature code. Journal of Chinese Information Processing, 2003,17(2):28−35 (in Chinese with English abstract).

[12] Manku GS, Jain A, Sarma AD. Detecting near duplicates for Web crawling. In: Proc. of the Int'l World Wide Web Conf. Committee (IW3C2). Banff, 2007. 141−149. [doi: 10.1145/1242572.1242592]

[13] Ye SZ, Wen JR, Ma WY. A systematic study on parameter correlations in large-scale duplicate document detection. In: Proc. of the 10th Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD). Singapore, 2006. 275−284. [doi: 10.1007/s10115-007-0071-9]

[14] Chowdhury A, Frieder O, Grossman D, McCabe MC. Collection statistics for fast duplicate document detection. ACM Trans. on Information System, 2002,20(2):171−191. [doi: 10.1145/506309.506311]

[15] Cooper JW, Coden AR, Brown EW. Detecting similar documents using salient terms. In: Proc. of the 11th ACM Int'l Conf. on Information and Knowledge Management (CIKM). Washington, 2002. 245−251. [doi: 10.1145/584792.584835]

[16] Conrad JG, Guo XS, Schriber CP. Online duplicate document detection: Signature reliability in a dynamic retrieval environment. In: Proc. of the 12th Int'l Conf. on Information and Knowledge Management (CIKM). New Orleans, 2003. 443−452.

[17] Kolcz A, Chowdhury A, Alspector J. Improved robustness of signature-based near-replicate detection via lexicon randomization. In: Proc. of the 10th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (SIGKDD). New York, 2004. 605−610. [doi: 10.1145/1014052.1014127]

[18] Watts DJ, Strogatz SH. Collective dynamics of small-world networks. Nature, 1998,393(6684):440−442. [doi: 10.1038/30918]

[19] Matsuo Y, Ohsawa Y, Ishizuka M. KeyWorld: Extracting keywords in a documents as a small world. In: Proc. of the 4th Int'l Conf. of Discovery Science. LNCS 2226, Washington, 2001. 271−281. [doi: 10.1007/3-540-45650-3_24]

[20] Cancho RF, Sole RV. The small world of human language. The Royal Society of London, Biological Sciences (Series B), 2001,268(1482):2261−2265. [doi: 10.1098/rspb.2001.1800]

[21] Zhu MX, Cai Z, Cai QS. Automatic keywords extraction of Chinese document using small world structure. In: Proc. of the Natural Language Processing and Knowledge Engineering Int'l Conf. 2003. 26−29. [doi: 10.1109/NLPKE.2003.1275946]

[22] Huang C, Tian YH, Zhou Z, Ling CX, Huang TJ. Keyphrase extraction using semantic networks structure analysis. In: Proc. of the IEEE Int'l Conf. on Data Mining (ICDM). Las Vegas, 2006. [doi: 10.1109/ICDM.2006.92]

[23] Cao YJ, Niu ZD, Dai LL, Zhao YM. Extracting of informative blocks from Web pages. In: Proc. of the 7th Int'l Conf. on Advanced Language Processing and Web Information Technology (ALPIT). Dalian, 2008. 544−549. [doi: 10.1109/ALPIT.2008.106]

[24] Hulth A. Combining machine learning and natural language processing for automatic keyword extraction [Ph.D. Thesis]. Department of Computer and Systems Sciences, Stockholm University, 2004.

[25] Lyon C, Nehaniv CL, Dickerson B. Entropy indictors for investigating early language process. In: Proc. of the EELC 2005. 2005. 143−150.

[26] Strogatz SH. Exploring complex networks. Nature, 2001,410(6825):268−276. [doi: 10.1038/35065725]

[27] Holyst JA, Sienkiewicz J, Fronczak A, Fronczak P, Suchecki K. Universal scaling of distances in complex networks. Physical Review E, 2005,72(2):026108. [doi: 10.1103%2fPhysRevE.72.026108]

[28] Hoad TC, Zobel J. Methods for identifying versioned and plagiarized documents. Journal of the American Society for Information Science and Technology, 2003,54(3):203−215.

附中文参考文献:

[1] 中国互联网络信息中心.第 16 次中国互联网络发展状况统计报告.2007. http://www.cnnic.net.cn/index/0E/00/11/index.htm

[5] 鲍军鹏,沈钧毅,刘晓东,宋擒豹.自然语言文档复制检测研究综述.软件学报,2003,14(10):1753−1760. http://www.jos.org.cn/1000-9825/14/1753.htm

[11] 吴平博,陈群秀.基于特征串的大规模中文网页快速去重算法研究.中文信息学报,2003,17(2):28−35.

曹玉娟(1973－),女,浙江宁波人,博士,高级工程师,主要研究领域为智能信息检索.

赵堃(1976－),男,博士,讲师,主要研究领域为 P2P 信息检索.

牛振东(1968－)男,博士,教授,博士生导师,主要研究领域为 Web 挖掘,数字图书馆.

彭学平(1979－),男,博士生,主要研究领域为个性化信息检索.