

文章编号:1001-9081(2010)06-1661-03

## 基于聚团词的大规模文本转载识别算法

张京阳<sup>1,3</sup>, 张华平<sup>2,3</sup>, 刘金刚<sup>1</sup>

(1. 首都师范大学 计算机科学联合研究院, 北京 100037;

2. 北京理工大学 计算机学院, 北京 100080; 3. 中国科学院 计算技术研究所, 北京 100190)

(zhangjy@golaxy.cn; zhanghp@software.ict.ac.cn)

**摘要:** 文本转载识别是指从大规模文本库中检测出内容相同或相近的文档集合, 在热门话题检测、搜索引擎结果凝练、学术文章抄袭识别等诸多应用上, 存在普遍的需求。为适应网络文本转载形式的日趋多样化, 并进一步提升实用系统效率, 对各种文本特征及比较算法进行了研究分析, 提出了基于聚团词的大规模文本转载识别算法, 即: 依据词语的分布属性, 识别并提取高分聚团词用于表征文本, 之后通过对文本集进行扩展线性比较与多维比较两次操作, 最终筛选出转载识别结果。对比实验表明: 该算法在准确率、召回率与效率上有较高的综合性能。

**关键词:** 转载识别; 聚团词; 特征选择; 扩展线性比较; 向量空间模型

**中图分类号:** TP391 **文献标志码:** A

### Large-scale document forward detection algorithm based on agglomerate-term

ZHANG Jing-yang<sup>1,3</sup>, ZHANG Hua-ping<sup>2,3</sup>, LIU Jin-gang<sup>1</sup>

(1. Joint Faculty of Computer Scientific Research, Capital Normal University, Beijing 100037, China;

2. School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100080, China;

3. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)

**Abstract:** Document forward detection is that to find out article collection of the same or close content from a large-scale text library. It has widespread demand in popular articles exploring, results organizing of search engine, copy detection and so on. To meet the growing diverse forms of Internet text forward and improve system efficiency, this paper discussed certain text features and researched some comparison algorithms. Then, the large-scale document forward detection algorithm based on agglomerate-term was introduced. Its principle is: first, detect and extract the agglomerate-term according to the term's distribution, and make it a key feature to characterize the text; then, set an extensive linear comparison and a multi-dimensional comparison on it; finally, compute the ultimate results of the forward detection. The experimental results show that the agglomerate-term algorithm has a better integrated performance of precision, recall and speed.

**Key words:** forward detection; Agglomerate-Term (AgT); feature selection; extensive linear comparison; Vector Space Model (VSM)

## 0 引言

自进入互联网时代起, 科技迅猛发展, 信息资源呈几何爆炸式增长。与此同时, 人们逐渐需要开始面对愈发严重的信息冗余问题。有资料统计, 在互联网上多达千亿的网页当中, 镜像网页所占比率高达 22%<sup>[1]</sup>。通过实验室对 2009 年 12 月 2—12 日 10 天内的境内外 100 家网络站点进行的新闻监测发现: 共采集到新闻文章 341 456 篇, 平均每篇文章转载 1.95 次, 转载数量的峰值接近 1 000, 每日热门新闻通常会被转载 100 次以上。信息冗余是互联网上信息管理不善的结果。互联网的开放特性使其在管理上难以像传统媒体那样, 能够在信息的发布与传播环节中进行把关控制。事实上, 只有少数网站真正拥有自己的专业编辑, 大部分群体会选择方便快捷的转载或是照抄的模式, 以达到个人目的。

针对这种现象, 学术界开始进行文本转载识别算法的研

究, 主要用于在海量文本中, 找出那些内容上相同或相近的文本集合。在多年的研究中, 不断有良好的算法涌现, 并成功构建出监控系统以应对各种需求。

文献[1-7]是针对早期文本转载识别需求, 分别从不同算法角度进行的尝试, 并成功实现于各种实用系统之上。但是, 在这些监控系统发展的同时, 文章转载形式也随之愈发多样化, 如段落乱序、词句修订、批量摘选等。而且网页编码具备高自由度, 从中定位并摘取的网页正文内容必然伴随噪声或者小规模形变, 这些都是需要考虑的问题。传统的解决方法若不加以修改重构, 已不再具备适用性。

近年来, 针对这些网络文本的变异, 各种研究又重新展开。文献[8]着重从计算语言学的角度入手, 力求改进识别效果。但一些较传统的基础模型严重限制住了系统的运行效率。而类似于文献[9]所述的研究, 完全是以分析超文本标记语言 (HTML) 的编程特性为优先, 这类方法只能应用于一

收稿日期: 2009-12-15; 修回日期: 2010-03-01。

基金项目: 国家 863 计划项目 (2007AA01Z438); 中国科学院计算技术研究所 2008 知识创新基金资助项目。

作者简介: 张京阳 (1985-), 男, 河北南宫人, 硕士研究生, 主要研究方向: 中文信息处理; 张华平 (1978-), 男, 江西波阳人, 副研究员, 博士, 主要研究方向: 舆情计算、计算语言学、中文信息处理; 刘金刚 (1963-), 男, 辽宁营口人, 教授, 博士, 主要研究方向: 智能接口。

小部分网站,满足一些定制性的需求,总体使用范围有限。

本文进一步寻求一种既能遵循语言学规律又可架构高效相似度比较算法的模型,从而使实用转载识别系统在效果与效率上能够达到一个更高的平衡点。

## 1 相关工作

在文本转载识别研究领域的发展进程中。经过一定的分析筛选,所有算法框架大致归为3类<sup>[10]</sup>,其中每种框架在细节的实现上又衍生出各式各样的解决方案。

框架1

$$(MD5( Abstract(P_i) )) = MD5( Abstract(P_j) )) \Rightarrow Mirror(P_i, P_j)$$

框架2

$$(MD5( Conc(sort(T_i)) )) = MD5( Conc(sort(T_j)) )) \Rightarrow Mirror(P_i, P_j)$$

框架3

$$MD5( Conc(sort(T_i)) ) = MD5( Conc(sort(T_j)) ) \left. \begin{array}{l} |W_i - W_j|^2 \\ |W_i|^2 + |W_j|^2 < \delta \end{array} \right\} \Rightarrow Mirror(P_i, P_j); 0 \leq \delta \leq 1$$

其中: $P_i$ 表示第*i*个网页,网页*i*的最高得分关键词top-*N*记为 $T_i = \{t_{i1}, t_{i2}, \dots, t_{in}\}$ ,对应的特征向量记为 $W_i = \{w_{i1}, w_{i2}, \dots, w_{in}\}$ ;  $Abstract(P_i)$ 表示网页摘要,  $Conc(T_i)$ 表示将 $T_i$ 向量进行简单字符串拼接,  $sort(T_i)$ 表示对向量做排序;  $MD5(X)$ 表示字符串*X*的哈希值;  $Mirror(P_i, P_j)$ 代表两网页判定近似,视为转载。

框架1是对网页摘要进行等值比较。这种方法的特点是运行效率极高,曾一度受到广泛关注。其变数在于网页摘要的获取方式,很多研究者从不同角度进行了尝试。典型的案例如文献[2],提出首先提取网页的正文,主要用于去除噪声,之后定位文本中某个特定标点(如句号、换行符等)的第*m*次出现,从该位置截取前后各5个字视为摘要。另一个案例,如知名搜索引擎网站vivisimo的聚类系统,主要基于动态摘要算法。用户输入检索需求后,系统找到与之匹配度最高的文本段视为摘要。框架1可避免对文本进行全扫描,在效率上已达到最高,但其处理范围有一定局限性,随转载形式的多样化发展,最终效果逐步下降。

框架2、3基于向量空间模型。其中框架2在两篇文档top-*N*关键词一致时便视为转载;框架3进一步增加在多维向量空间上的相似度计算,综合考虑了文章的结构属性。相应案例如文献[3],拼合文档中句子的首字与尾字视为关键词;另有SCAM(Stanford Copy Analysis Method)原型系统,以词频为统计元素,高频词即为关键词。类似案例还有很多,不同方法各有不同的性能展现。

## 2 基于聚团词的大规模文本转载识别算法

转载识别系统的主要工作流程一般情况如图1所示。

### 2.1 聚团词提取

通常,一篇文章不会通篇参与计算,而是借由分析文章内容的中心思想与作者的写作手法,提取出一组能够大致表征

文章的关键词序列用于计算。该关键词序列对应的提取过程即称之为特征选择。作为文本比较的基准,文本特征应更加关注于对整篇文章或是文章片断的区分程度。传统算法中多以词频-逆文档频(TF-IDF)概念计算特征串的权重。但是在实际应用中,TF-IDF对文档的区分程度有限,有时又容易出现多个备选词得分近似的情况,无法找到正确的取舍方式,最终影响识别效果。并且,诸如IDF这类全局性属性,需要在对整个文档集进行一定程度的统计之后才能够估算得到,这也限制了系统的运行效率。从经验上讲,在筛选关键词时若能兼顾考虑文章的语义结构,识别正确率可显著提升。但是,分析语义结构至少需要对文章进行一次完整的扫描,这个步骤将会是系统整套识别流程的主要时间消耗。

网页/原始文本

粗过滤/正文抽取

关键词提取

扩展线性比较

多维比较

识别结果

图1 转载识别系统的主要工作流程

聚团词(Agglomerate-Term, AgT)是指在文本中出现了聚团现象的词。本研究定义,在下面两种条件下,词的聚团属性得分给予增加:

1) 某个词在小规模文本片断中多次出现。很明显该条件亦符合局部性原理,使得某个词即使在全文中不具备高频特性,也有机会跻身关键词行列。

2) 某个词在文本中出现多次聚团现象。这里做了一个假设,认为一个词在整篇文本中的分布越是不均匀,其重要性越高。在该假设条件下,一些常用的领域词有可能会被看作垃圾串而被过滤掉。实验表明,这种附带效果亦起到了一定的积极作用。

聚团词提取算法如下:

$$score(W) = \sum_i LS(W, i) \quad (1)$$

$$LS(W, i) = f(LS(W, i-1), \Delta D(W, i)) \quad (2)$$

$$f(a, b) = (a + m(b)) \cdot k(b) \quad (3)$$

其中: $score(W)$ 表示词*W*的总得分,  $LS(W, i)$ 为*W*在第*i*次出现的单次得分,  $\Delta D(W, i)$ 表示*W*相邻两次出现的距离;  $f$ 为评价函数,首要参数*k*的取值要保证*f*关于自变量*b*单调递减,次要参数*m*的取值要保证*f*关于*b*连续,从而体现出单词聚簇程度。参数函数*m*与*k*建议通过机器学习或遗传算法予以确定,本系统暂时使用经验函数代替。

由式(1)~(3)可知,所有的词在文章中的每一次出现都将被计算一个得分,而该得分仅与这个词在上一个出现点的得分以及两次出现的位置跨度有关。最终,将同一个词的得分累加,作为该词的总得分。这种计算方式,既保证了算法的效率,又可避免某两个词之间得分一致,难以取舍的局面。

### 2.2 扩展线性比较

众所周知,多维数据运算在大规模应用中运算量巨大,系

统必然要在该操作之前进行过滤预处理。线性过滤算法以高效著称,但误删率太高,严重影响到最终结果的召回效果。目前仍未找到一个合适的算法,能在对文档集进行高效过滤的同时,保持一个较小的误删率。

本系统由此使用扩展线性比较算法,即同时做多次相互不覆盖的线性预处理操作,之后将多个过滤结果进行合并。

首先将前面提取的关键词,依权重降序序列摘选 top- $N$ ,对选取的多个关键词做字符串排序后进行简单连接。之后,将得到的词串计算哈希值用于比较。其中字符串排序操作是为了忽略多个关键词之间的位置关系,以增加召回文章数量。这里依据  $N$  的不同取值作为不同线性过滤的划分,当  $N$  取 3 ~ 5 时,过滤效果与召回效果较为平衡<sup>[10]</sup>。最终,系统合并 3 种线性过滤的结果,作为下一个步骤的输入。

例 文章  $P_1$ - $P_4$  对应的关键词序列  $T_1 \sim T_4$  如下:

$$T_1 = \{b, c, d, f, a\} = \{abcdf\}_5 = \{bcd\}_4 = \{bcd\}_3$$

$$T_2 = \{b, c, d, e, a\} = \{abcde\}_5 = \{bcde\}_4 = \{bcd\}_3$$

$$T_3 = \{a, b, c, d, e\} = \{abcde\}_5 = \{abcd\}_4 = \{abc\}_3$$

$$T_4 = \{a, b, e, f, c\} = \{abcef\}_5 = \{abef\}_4 = \{abe\}_3$$

第 1 步 独立计算  $N$  不同取值的线性比较:

$$N = 3 \text{ 时, } \{(T_1 T_2) \dots\};$$

$$N = 4 \text{ 时, } \{\emptyset\};$$

$$N = 5 \text{ 时, } \{(T_2 T_3) \dots\}.$$

第 2 步 将各次线性比较结果予以合并,得到过滤划分

$$\{(T_1 T_2 T_3) \dots\}.$$

### 2.3 多维比较

经过线性过滤之后,计算数据集不再是一个巨大的向量集合,取代的是若干极小型的向量集合,放置于向量空间模型 (Vector Space Model, VSM) 之中时,可以大幅度减少计算代价。在这里本文不继续使用词的聚团属性作为计算基准,转用词频 (TF) 属性代替。原因是在本系统中,词的聚团得分在计算时,引入了过剩的词语级别区分度,这样会使得计算向量相似度时的判定阈值难以把握。

文本相似度的计算公式如下:

$$\text{sim}(\mathbf{W}_i, \mathbf{W}_j) = \frac{\mathbf{W}_i \cdot \mathbf{W}_j}{\|\mathbf{W}_i\| \times \|\mathbf{W}_j\|} = \frac{\sum_k w_{ik} \cdot w_{jk}}{\sqrt{\sum_k w_{ik}^2 \cdot \sum_k w_{jk}^2}}$$

$$w_i = \text{tf}(\mathbf{W}_i) / \text{len}(\mathbf{P}_i)$$

其中:  $\text{sim}(\mathbf{W}_i, \mathbf{W}_j)$  表示两个关键词向量的相似系数,  $\text{tf}(w)$  代表词频,  $\text{len}(P)$  代表文档长度。

## 3 实验结果及分析

### 3.1 实验环境与实验数据

机器配置:PC 机双核 CPU 2.01 GHz, 内存 1.93 GB。数据源选用当前市场上代表性商用转载识别系统的测试文档集 (出自海量信息技术有限公司)。本文实验将对基于聚团词的大规模文本转载识别算法、基于 TF-IDF 的转载识别算法、基于标点符号的大规模网页快速去重算法和代表性商用系统

算法 4 种文本转载识别算法做横向比较。

### 3.2 实验结果

通过对本算法的运行时效率分析,可以看到在向量空间模型上的最后一个比较环节是高效的,主要时间花费于提取聚团词时对文档集进行的一次完整扫描上。值得关注的是,在本算法中,使用的是静态关键词获取方式,关键词提取操作可以与后续操作分离,并可作为文本指纹存储于外部设备。这种方式亦符合几类大规模数据处理流程。

表 1 AgT 关键词提取效果采样

文本标题	关键词序列
中国软件与 IT 服务市场增势趋缓	服务 市场 中国 上半年 软件 行业...
半导体业合作开发已成大势	曙光 公司 掩膜 技术 制造商 半导体...
中国移动市场 1 月份市场分析报告	联通 彩信 中国移动 短信 手机 用户...
2002—2003 年中国家用台式 PC 市场概况	服务 广告 电脑 互联网 家用 渠道...
主流厂商 MSTP 产品分析	业务 支持 符合 网络 国家标准 国际...
专注与创新是半导体工业的生命线	支持 电源 产品 芯片 调试 充电器...

表 2 AgT 效率分析

文档集大小	耗时			s
	关键词提取	线性比较	多维比较	
2900	4.500	0.014	0.001	4.515
9825	8.328	0.003	0.016	8.437

实验中为避免 TF-IDF 运行时间过长,向量空间维度取值为 10。TF-IDF 系统相似度阈值设定为 0.95。

表 3 各算法性能对比

文档集大小	算法	运行时间/s	召回率/%	准确率/%
2900	聚团词算法	4.515	90.2	96.5
	TF-IDF	36.061	85.0	88.1
	标点符号法	1.559	58.6	71.3
	某商用系统	29.674	56.1	100
9852	聚团词算法	8.437	89.8	97.0
	TF-IDF	380.934	91.2	81.3
	标点符号法	2.372	63.1	69.9
	某商用系统	51.441	68.3	100

在对 4 种算法的性能对比中,可以看出本算法在识别性能各方面具有高度的平衡。准确率高,运行时间仅次于无需对文本进行全扫描的标点符号法。尤其值得关注的是,本算法召回效果较之其他各算法有非常显著的提升。

## 4 结语

本文引入聚团词概念作为文本特征,以应对转载形式日趋多样化的大规模文本转载识别要求。该特征能够在一定程度上对文本语义进行描述,并能够配合扩展线性比较与多维比较处理模型,高效地完成任务,使文本转载识别系统得到了进一步的优化,能有效应对当下需求。

计划下一步工作主要从两方面进行:

1) 应用遗传算法将系统参数进行调优;

2) 在大规模语料实验统计的基础上优化分词词典,进一步提高系统性能。

参考文献:

- [1] SHIVAKUMAR N, GARCIA M H. Finding near-replicas of documents on the Web[C]// WebDB'98. Berlin Heidelberg: Springer-Verlag, 1999: 204-212.

(下转第 1670 页)



中各节点的内在信息量之和  $IC(V_{a \cup b})$  的比值作为节点  $v_a$  和  $v_b$  代表的术语之间的语义相似度,用公式表示为:

$$sim(v_a, v_b) = \frac{IC(V_{a \cap b})}{IC(V_{a \cup b})} \quad (14)$$

根据图1所示,采用基于有向无环图和内在信息量的混合方法计算节点  $f$  和  $g$  代表的术语的语义相似度。图中所有节点的内在信息量已经标注,  $f$  的子图和  $g$  的子图的交集包括  $\{a, b, c\}$ , 并集包括  $\{a, b, c, d, f, g\}$ 。交集中节点的内在信息量之和是 0.394, 并集中节点的内在信息量之和是 2.866。最后得到节点  $f$  和  $g$  代表的术语的语义相似度是 0.137。

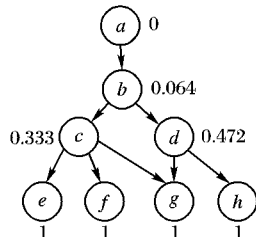


图1 术语所在的有向无环图

### 3 实验与分析

评价语义相似度方法,目前被大多数研究者使用的是 Rubenstein and Goodenough 提出的 65 对单词对以及给出的相似度标准。Rubenstein and Goodenough 让 51 个测试对象根据语义的相似度对 65 对单词进行打分,分值为 0~4,语义相似度越高的单词对分值越高。本文将基于有向无环图和内在信息量的混合方法和其他方法分别计算 65 对单词对的语义相似度,和人工判断对比,通过式(15)得到 Pearson 相关系数:

$$r = \frac{\sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n}}{\sqrt{\left( \sum_{i=1}^n x_i^2 - \frac{\left( \sum_{i=1}^n x_i \right)^2}{n} \right) \left( \sum_{i=1}^n y_i^2 - \frac{\left( \sum_{i=1}^n y_i \right)^2}{n} \right)}} \quad (15)$$

其中:  $r$  表示相关系数,  $x_i$  表示计算方法得到的第  $i$  对单词的语义相似度,  $y_i$  表示人工判断得到的第  $i$  对单词的语义相似度,  $n$  表示单词对的数量。实验结果如表1所示。

实验结果表明,本文方法有比较高的准确度。

### 4 结语

本文提出的基于有向无环图和内在信息量的混合方法采用基于内在信息量的计算方法避免了分析语料库的问题,同

时综合考虑了术语所在的有向无环图的结构,使得语义相似度的计算更符合人工的判断。该方法可以应用于和 WordNet 类似的本体的术语间语义相似度的计算。

表1 各方法的相关系数

语义相似度计算方法	相关系数
Resnik 方法	0.846
Lin 方法	0.858
Jiang 方法	0.864
Zhong 方法	0.754
Li Yuhua 方法	0.908
OSS 方法	0.911
本文方法	0.914

### 参考文献:

- [1] RESNIK P. Using information content to evaluate semantic similarity in a taxonomy [C]// Proceedings of the 14th International Joint Conference on Artificial Intelligence. California: Morgan Kaufmann Publishers, 1995: 448-453.
- [2] LIN D. An information-theoretic definition of similarity [C]// Proceedings of 15th International Conference on Machine Learning. California: Morgan Kaufmann Publishers, 1998: 296-304.
- [3] JIANG J, CONRATH D. Semantic similarity based on corpus statistics and lexical taxonomy [C]// Proceedings of International Conference on Research in Computational Linguistics. Washington, DC: IEEE, 1997: 19-33.
- [4] NORMAN L COO, GARNER B, TSUI E, et al. Semantic distance in conceptual graphs [C]// Proceedings of Fourth Annual Workshop on Conceptual Structures. New York: Ellis Horwood, 1992: 149-154.
- [5] ZHONG J W, ZHU H P, LI J M, et al. Conceptual graph matching for semantic search [C]// Proceedings of the 10th International Conference on Conceptual Structure. Berlin: Springer-Verlag, 2002: 92-106.
- [6] LI Y, BANDAR A, MCLEAN D. An approach for measuring semantic similarity between words using multiple information sources [J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(4): 871-882.
- [7] ZUBER V S, FALTINGS B. OSS: A semantic similarity function based on hierarchical ontologies [C]// Proceedings of the 20th International Joint Conference on Artificial Intelligence. California: Morgan Kaufmann Publishers, 2007: 551-556.
- [8] SECO N, VEALE T, HAYES J. An intrinsic information content metric for semantic similarity in WordNet [C]// Proceedings of the 16th European Conference on Artificial Intelligence. Amsterdam: IOS Press, 2004: 1089-1090.

(上接第1663页)

- [2] 张刚, 刘挺, 郑实福, 等. 大规模网页快速去重算法 [C]// 中文信息学会二十周年学术会议论文集: 续集. 北京: 清华大学出版社, 2001: 18-25.
- [3] 吴平博, 陈群秀, 马亮. 基于特征串的大规模中文网页快速去重算法研究 [J]. 中文信息学报, 2003, 17(2): 28-35.
- [4] 孔素然, 黄萱菁. 基于模糊匹配思想的网页去重算法 [D]. 上海: 复旦大学, 2006.
- [5] 庞剑锋, 卜东波, 白硕. 基于向量空间模型的文本自动分类系统的研究与实现 [J]. 计算机应用研究, 2001, 18(9): 23-26.
- [6] 王哲. 基于特征码的网页去重算法研究 [J]. 山东广播电视大学

学报, 2009(1): 14-16.

- [7] 李卫, 刘建毅, 王枫. 基于全信息的网络文本信息去重算法研究 [C]// 第十一届中国人工智能学术年会. 北京: 北京邮电大学出版社, 2005: 1276-1281.
- [8] 连浩, 刘悦, 许洪波, 等. 改进的基于布尔模型的网页查重算法 [J]. 计算机应用研究, 2007, 24(2): 36-39.
- [9] 韩冰, 林鸿飞. 大规模文本去重策略研究 [D]. 大连: 大连理工大学, 2008.
- [10] 王建勇, 谢正茂, 雷鸣, 等. 近似镜像网页检测算法的研究与评价 [J]. 电子学报, 2000, 28(21): 129-132.