

中文农业网页去重及相似度判断研究

赵 涛¹, 张太红^{1,2}, 陈燕红¹

(1. 新疆农业大学 计算机与信息工程学院, 新疆 乌鲁木齐 830052;

2. 中国农业大学 信息与电气工程学院, 北京 100083)

摘 要:随着信息技术的飞速发展,互联网中的网页急剧增长,在这海量、繁杂的网页中却呈现出一定比例的重复网页及近似网页。为了减少农业领域中近似及重复网页对农业垂直搜索引擎性能的影响,文中首先使用 MD5 算法去除网页集合中完全相同的网页,再利用向量空间模型(VSM)、基于知网的语义相似度模型及潜在语义分析(LSA)三种相似度判断方法对其余网页的相似度进行计算。实验结果显示,当相似度阈值 $r=60\%$ 、维数 $K=250$ 时,潜在语义分析(LSA)的综合评价 F_1 测度最高,且准确率达到了 90.5%。

关键词:中文农业网页;MD5;向量空间模型;知网;潜在语义分析

中图分类号:TP393

文献标识码:A

文章编号:1673-629X(2015)01-0191-04

doi:10.3969/j.issn.1673-629X.2015.01.043

Research on Duplicate Removal and Similarity Evaluation of Chinese Agricultural Web Pages

ZHAO Tao¹, ZHANG Tai-hong^{1,2}, CHEN Yan-hong¹

(1. College of Computer & Information Engineering, Xinjiang Agricultural University,

Urumqi 830052, China;

2. College of Information and Electrical Engineering, China Agricultural University,

Beijing 100083, China)

Abstract: With the rapid development of information technology, the Internet Web pages are growing sharply. In this massive, complex pages, preach a certain percentage of duplicate pages and similar pages. In order to reduce the influence of agricultural field approximation and repeated Web pages on agricultural vertical search engine performance, first use the MD5 algorithm to remove the same Web pages in the Web page set, then through three kinds of methods which include the Vector Space Model (VSM), semantic similarity model based on HowNet and Latent Semantic Analysis (LSA), calculate the similarity of the rest Web pages. The experimental results show that when the similarity threshold is 60% ($r=60\%$), the dimension is 250 ($K=250$), the F_1 comprehensive evaluation measure of LSA is highest, and the accuracy rate has reached 90.5%.

Key words: Chinese agricultural Web page; MD5; vector space model; HowNet; latent semantic analysis

0 引言

截止到2013年4月,中国农业网站数量已超过4万个,以平均每个网站5000张网页计算,农业网页数量以达到2亿^[1]。数量庞大的农业网页中存在着一定数量的重复或近似重复的网页,它们大多来自镜像网站和网页的转载,研究表明Internet上内容近似的网页占30%~45%^[2]。这些冗余的信息需要额外的存储资源,也会降低索引效率,直接影响搜索引擎的整体性

能。因此如何对这些农业网页中重复以及近似重复的网页进行有效的管理成为农业垂直搜索引擎领域研究的课题之一。

1 中文农业网页去重及相似度计算系统框架

文中从农业网站中获取网页样本,使用MD5算法去除完全重复的网页,对于剩余网页结合HtmlParser

收稿日期:2014-03-10

修回日期:2014-06-12

网络出版时间:2014-10-23

基金项目:新疆自治区高校科研计划项目(XJEDU2013S13);新疆农业大学前期资助课题(XJAU201117)

作者简介:赵 涛(1989-),女,新疆焉耆人,硕士,研究方向为数据库;张太红,教授,主要从事数据库技术等方面的研究。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20141023.1520.043.html>

提取出正文内容,对中文文本分词,过滤停用词、英文字符及英文标点符号,建立向量空间模型(VSM)、基于知网的语义相似度模型、潜在语义分析(LSA)模型分别计算相似度。中文农业网页去重及相似度判断模型的算法框架见图1(判别方法包括上面提到的三种方法)。

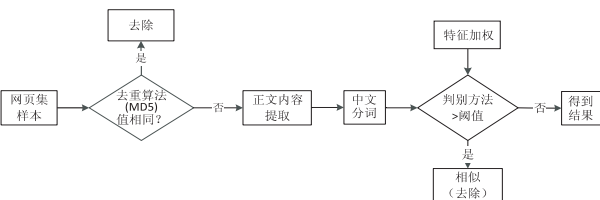


图1 网页去重及相似度判断算法框架图

2 主要技术简介

2.1 MD5 算法(信息-摘要算法)

MD5 的作用是让大容量信息在用数字签名软件签署私人密钥前被“压缩”成一种保密的格式(就是把一个任意长度的字节串变换成一定长的大整数)。对 MD5 算法简要的叙述为:MD5 以 512 位分组来处理输入的信息,且每一分组又被划分为 16 个 32 位子分组,经过了一系列的处理后,算法的输出由 4 个 32 位分组组成,将这 4 个 32 位分组级联后将生成一个 128 位散列值^[3]。文中即是使用 MD5 算法将原始农业网页数据经过处理得到一个 128 位的 MD5 编码,通过比较 MD5 编码消除相同的网页,降低了下一步相似度判断中的计算量,提高了效率。

2.2 正文内容提取及预处理

网页中包含大量的噪声,如广告、导航条、链接、JavaScript 脚本等内容,网页所要表达的主要信息被隐藏在无关的内容和结构中。因此需要提取网页的正文内容。对于网页中存在的大量图片及链接页面,这里不做讨论。一般网页的内容都包含在 table、tr、td、div、p、I、B、Strong 等特征标签中^[4]。文中采用 Html-Parser 解析工具将网页解析为 DOM 标签树结构,调用 parser 的 elements 方法获取所有的标签对应的 node,然后使用递归遍历所有满足要求的 node 节点。这样做的好处是处理降噪转变为树节点的遍历和提取有效信息的过程。但是对于网页内容的提取,不同网站的设计,对正文部分没有一个统一的规则,所以误差是避免不了的。

2.3 中文分词

对于文档相似度计算,分词是基础和关键,采用高效的分词算法能够极大地提高网页相似度计算结果的准确性。文献[5]对 paoding、ik、imdict、mmseg4j 四种分词进行了比较,结果表示 paoding 在分词效果、速度、扩展性性能都比其他三种效果好;文献[6]中对 je、ik、

paoding、中科院的 ictclas 进行实验对比,实验显示 paoding 在分词效果、性能、准确率方面都表现最好。所以文中在实验过程中使用 paoding 分词器。

2.4 向量空间模型(VSM)

向量空间模型是近年使用较广泛的信息检索模型,运行效率高,简单,并且取得了很好的成果。其原理是将一个文本的处理简化为向量空间中的向量运算,把文本转换成计算机能够识别的向量,一个词语对应向量中的一个维度,从而简化了文本中关键词之间的复杂关系。通过向量的方式来计算出文本之间的相似度^[7-8]。而网页中的内容也可以看作是文本,在网页中提取正文,就可以计算不同网页中文本的相似度。

目前在信息处理方向上,向量空间模型的基本思想是以向量来表示文档: $[w_1, w_2, \dots, w_n]$, w_j 表示第 j 个特征值的权重,使用最广泛的权重计算方法 TF-IDF (Term Frequency Inverse Document Frequency),计算词语在单个文本中出现的频度与词语在整个文本集合中出现的频度,即特征词 t 在文档 d_i 中出现频率越高,词语 t 越重要;特征词 t 在文档集合中的文档频率 $df(t)$ 越高,词语 t 越不重要。公式表示如下:

$$W_{ij} = tf_{ij} \times idf(t_j) = tf_{ij} \times \log\left(\frac{N}{df(t_j)} + 0.01\right) \quad (1)$$

其中, tf_{ij} 表示 t_j 在文档 d_i 中的词频; $idf(t_j)$ 表示特征词 t_j 在样本集中的文档频率; N 为样本集的总数。

$$Sim(d_i, d_j) = \frac{\sum_{a=1}^n w_{ia} * w_{ja}}{\sqrt{\left(\sum_{a=1}^n w_{ia}^2\right) \left(\sum_{a=1}^n w_{ja}^2\right)}} \quad (2)$$

其中, d_i 、 d_j 表示文档 i 和文档 j ; n 表示特征向量的维数。

2.5 基于知网的语义相似度计算

2.5.1 知网的介绍

知网(HowNet)是一个常识知识库,它描述的对象是中文和英文的词语所代表的概念,揭示两个概念属性之间的关系^[9]。

“概念”和“义原”是知网中两个非常重要的概念,“概念”用来描述词语语义,每个词语可以用多个“概念”来描述^[10];“义原”是最基本的单位,是一种“知识表示语言”,是用来对“概念”进行描述的一种词语,是描述“概念”的最基本单位。知网中共包含 1 500 多个义原。

知网体系中,一般用一组“义原”组合来表示一个“概念”,“义原”是层次体系中的一个节点。“义原”之间有八种关系:上下位关系、语义相同关系、语义相反关系、语义相对关系、寄主和属性关系、部件和整体关系、成品和材料关系、角色和事件关系^[11]。

2.5.2 文本语义相似度的计算

知网基本上由两类词组成,实词和虚词。实词的表达方式比较复杂,由一系列“语义描述式”组成实词的表达式^[12],有三种“语义描述式”:独立义原描述式,其中独立义原分两种,第一独立义原和其他独立义原,关系义原描述式,符号义原描述式。对概念的各个部分的相似度分别记为 $\text{Sim}_1(S_1, S_2)$ 、 $\text{Sim}_2(S_1, S_2)$ 、 $\text{Sim}_3(S_1, S_2)$ 、 $\text{Sim}_4(S_1, S_2)$,则概念的整体相似度记为:

$$\text{Sim}(C_1, C_2) = \sum_{i=1}^4 \beta_i \text{Sim}_i(S_1, S_2) \tag{3}$$

其中, $\beta_i (1 \leq i \leq 4)$ 是人为可以变动的参数,根据实际需要,改变大小,得到符合实际情况的结果,且有 $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1, \beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4$ 。

通过词语的相似度直接计算整个篇章的同类词性的词语集合,另外为了防止两个文档长度不一而造成词语的空匹配,采用了双向最大匹配,反之亦然,然后取两个相似度的平均值^[12]。其计算公式如下:

$$\text{Sim}(T_1, T_2) = \frac{1}{2} \left(\frac{\sum_{s \in |T_1|} (\max \text{Sim}(s, T_2) * \text{idf}(s))}{\sum_{s \in |T_1|} \text{idf}(s)} + \frac{\sum_{s \in |T_2|} (\max \text{Sim}(s, T_1) * \text{idf}(s))}{\sum_{s \in |T_2|} \text{idf}(s)} \right) \tag{4}$$

2.6 潜在语义分析(LSA)

潜在语义分析即 LSA (Latent Semantic Analysis), 是一种信息检索代数模型,是用于知识获取和展示的计算理论和方法。它将高维的特征空间映射到低维的潜在语义空间,从而消除了原始向量空间中的一些“噪声”,使向量空间维度降低,提高了信息检索的精确度^[13]。

LSA 具体实现:

建立一个词语对文档的语义空间,即构建 $M \times N$ 的矩阵 $C = [C_{ij}]_{M \times N}$ (其中每一行代表特征词 i 在所有文本中的特征权值向量, C_{ij} 为特征项 i 在文本 j 中的权值)。由于每个特征词只在少量的文本中出现,故矩阵 C 为高维稀疏矩阵,于是需要对 C 进行 SVD 分解。分解后的矩阵 C 被分解成 3 个矩阵的乘积^[14],即:

$$C = U \Sigma V^T \tag{5}$$

其中,矩阵 Σ 的对角线上的元素等于 C 的奇异值; U 和 V 的列分别是奇异值中的左、右奇异向量,这里 V^T 为 V 的转置。

LSA 在矩阵 C 分解成 $C = U \Sigma V^T$ 三个矩阵乘积的基础上(设 $M > N, \text{rank}(C) = r$,存在 $K, K < r$ 且 $K \ll \min(m, n)$),从 Σ 中选取 K 个最大的奇异值舍弃其他的奇异值使矩阵降维。那么矩阵 Σ 就变成 $K \times K$ 的奇异值对角矩阵 Σ_K, U_K 为 U 去掉 $M - K$ 个列向量的正交

矩阵, V_K 为 V 去掉 $N - K$ 个列向量的正交矩阵,通过奇异值分解的反运算得到矩阵 C 的近似矩阵。

$$C \approx C_K = U_K \Sigma_K V_K^T \tag{6}$$

余弦相似法是常用的相似度计算模型,两个向量的夹角越大,它们的相似度越小;相反,如果夹角越小,它们的相似度越大。这里以 C_K 的转置 C_K^T 为依托,使用式(2)计算任意 2 个文本向量之间的相似度。

3 实验分析及结果

3.1 评价标准及方法

通常在文本相似度计算中是通过准确率、召回率、 F_1 测度三种评价标准来评价实验结果。本实验首先建立一个相似网页集合,共 1 110 篇网页,再由人工鉴别,将相似或重复的网页放在一组中,共鉴别出 225 组网页集,且每组网页集中大概有 2 ~ 14 张近似重复网页。文中主要结合以上三种评价标准及三种判别方法对结果进行测试。计算公式中采用文献[4]中设定的参数,参数为:

$$\text{准确率 (Precision)} = \frac{\text{正确结果的数量}}{\text{所有返回结果的数量}} \tag{7}$$

$$\text{召回率 (Recall)} = \frac{\text{实际识别出的正确结果}}{\text{总的正确结果}} \tag{8}$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

判别方法通过编码转换,转换成统一的编码规则,对网页集采用 MD5 算法去除完全重复的网页,再提取出网页的正文内容,结合庖丁解牛 (paoding) 分词器进行分词,在此基础上分别使用三种方法相对应的公式计算相似度。

向量空间模型:使用向量空间模型的余弦公式(2)计算网页的相似度。

基于知网的语义相似度模型:首先使用公式(3)计算出词语的相似度,再使用公式(4)计算网页的相似度。

潜在语义分析:构造 term-document 矩阵 C ,使用式(5)对矩阵 C 进行奇异值分解,然后通过降维使用式(6)得到 C_K ,并通过公式(2)计算网页的相似度。

3.2 实验结果

本实验首先通过 MD5 方法去除了 41 篇与其他网页完全重复的网页,剩余 1 069 篇网页利用以上评测方法分别取不同的相似度阈值 ($r = 90\%$ 、 $r = 80\%$ 、 $r = 70\%$ 、 $r = 60\%$ 、 $r = 50\%$ 、 $r = 40\%$) 来对测试网页进行比较。实验结果如下:

基于语料库的实验结果是知网的结果优于向量空间模型的结果,但在文中却相反,主要原因在于本测试集中有很多农业领域的词汇,而知网在计算词语相似

度时如果遇到了没有录用的词则相似度为0,再结合词的权重,使得整篇文章与其他文章的相似度降低。从表1和表2两张表中可以看出,当相似度阈值 $r=60\%$ 时,VSM的结果(F_1 测度=88.1%)和知网的结果(F_1 测度=86.1%)都表现最优,且VSM的结果好于知网的结果。

表1 VSM中不同阈值下的评价方法结果

阈值	准确率	召回率	F 测度
90%	0.946	0.753	0.839
80%	0.940	0.803	0.866
70%	0.933	0.817	0.872
60%	0.909	0.856	0.881
50%	0.864	0.885	0.861
40%	0.828	0.896	0.861

表2 HowNet中不同阈值下的评价方法结果

阈值	准确率	召回率	F 测度
90%	0.922	0.733	0.817
80%	0.905	0.773	0.834
70%	0.884	0.803	0.841
60%	0.878	0.844	0.861
50%	0.841	0.847	0.844
40%	0.786	0.852	0.818

但在使用LSA算法时,需要选取合适的 K 值, K 值一般在100~300之间选取^[11]。文中选择的 K 值为300,250,200,150,100。

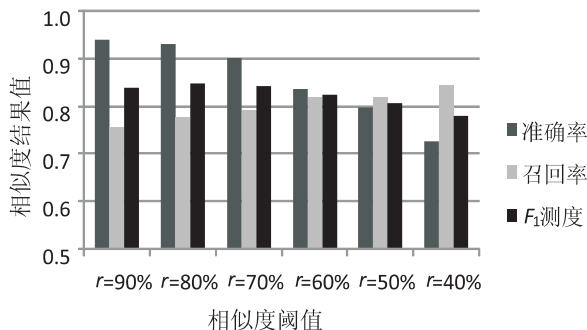


图2 LSA中 K 值=300时的结果

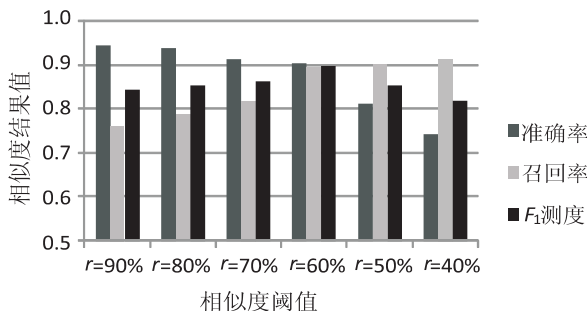


图3 LSA中 K 值=250时的结果

从图2~图6中可以看出, K 值的选取直接影响着降维效果。选取的 K 值应该足够小,去除不该保留的

“噪声”,又要足够大,以保留潜在语义结构中的主要框架, K 值的过大或过小都会给原始语义造成影响。根据实验结果得知,随着相似度阈值的越来越小,准确率越来越低,召回率越来越高,且根据多组实验的综合评测 F_1 测度值显示,当 $K=250$,相似度阈值 $r=60\%$ 时, F_1 测度最高,且准确率达到90.5%。再将此结果与VSM的结果对比,LSA的结果优于VSM的结果。

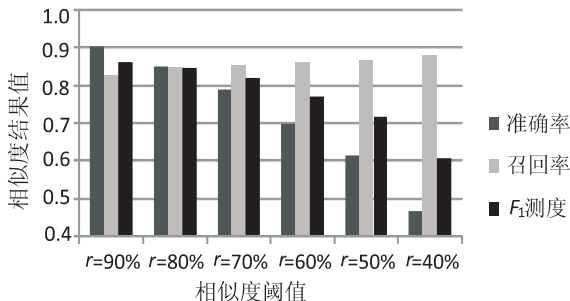


图4 LSA中 K 值=200时的结果

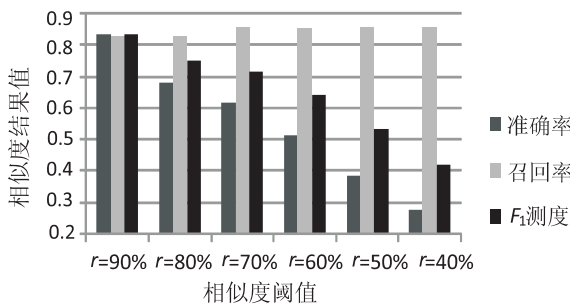


图5 LSA中 K 值=150时的结果

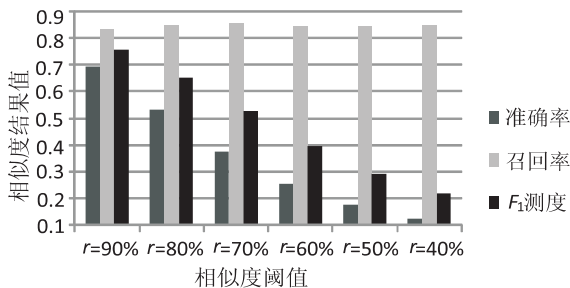


图6 LSA中 K 值=100时的结果

4 结束语

文中基于农业网页中存在大量近似及重复网页的特点,重点研究了MD5算法及向量空间模型、基于知网的语义相似度计算、潜在语义分析三种相似度判断算法,并对相似度结果进行了综合性的对比和分析。这对网页进行检索、比较不同网页内容、对网页进行查重以及对网页内容进行查询筛选等奠定了良好的基础。但LSA的缺点在于它的效果依赖于上下文信息及计算量过大,可能计算效率不是很高。

在以后的研究中,将对现有算法进行进一步的改进,提高相似度算法精度及算法效率,以取得更好的相似度判断效果。

3 结束语

对于任意波形的测量来说,所得测量结果的准确度、精确性以及测量速率和很多因素有关。在实验方案中,不仅和器件性能的优良性、先进性以及误差特性等有关,还和系统中待测脉冲与门脉冲之间的相对时延和测量速率有关。当待测脉冲频率恒定,增大可变时延的抽样密度,相应的测量速率会降低。而当加快测量速率时,抽样密度则被迫下降,导致待测脉冲细节丢失,测量结果失真。所以此测量系统的准确性,与可变时延的抽样密度成正比,同时和测量速率之间成反比。因此,一个好的测量装置依赖于两者之间的有效平衡。

参考文献:

- [1] Fork R L, Greene B I, Shank C V. Generation of optical pulses shorter than 0.1 psec by colliding pulse mode locking[J]. Applied Physics Letters, 1981, 38(9): 671-672.
- [2] Spence D E, Kean P N, Sibbett W. 60-fsec pulse generation from a self-mode-locked Ti:sapphire laser[J]. Optics Letters, 1991, 16(1): 42-44.
- [3] Khan M H, Shen Hao, Xuan Yi, et al. Ultrabroad-bandwidth arbitrary radiofrequency waveform generation with a silicon photonic chip-based spectral shaper[J]. Nature Photonics, 2010, 4: 117-122.
- [4] Wang Chao, Yao Jianping. Large time-bandwidth product microwave arbitrary waveform generation using a spatially dis-

crete chirped fiber bragg grating[J]. Journal of Lightwave Technology, 2010, 28(11): 1652-1660.

- [5] 文锦辉, 刘俊, 张慧, 等. 改进型零附加相位光谱相位相干电场重构系统对啁啾脉冲的测量[J]. 物理学报, 2010, 59(1): 370-375.
- [6] 张慧, 卢娟, 文锦辉, 等. 不同波长飞秒脉冲的相位测量[J]. 物理学报, 2011, 60(12): 270-276.
- [7] 庞杰. 基于微结构光纤的超短脉冲测量[D]. 北京: 北京邮电大学, 2012.
- [8] 刘佳. 飞秒激光成丝相互作用及诊断方法[D]. 武汉: 华东师范大学, 2013.
- [9] 龙井华, 高继华, 巨养锋, 等. 用 SHG-FROG 方法测量超短光脉冲的振幅和相位[J]. 光子学报, 2002, 31(10): 1292-1296.
- [10] 文汝红. 在频率分辨光学门法中用矩阵方法回归超短脉冲[J]. 宜春学院学报, 2008, 30(6): 4-6.
- [11] 王亚平, 吴重庆, 杨双收. 超短光脉冲的测量技术[J]. 中国科学信息, 2007(5): 279-280.
- [12] 张建忠. 超短光脉冲的频率分辨光学开关法测量研究[J]. 激光技术, 2008, 32(2): 194-197.
- [13] Jiang Zhi, Leaird D E, Weiner A M. Width and wavelength tunable optical RZ pulse generation and RZ-to-NRZ format conversion at 10 GHz using spectral line-by-line control[J]. IEEE Photonics Technology Letters, 2005, 17(12): 2733-2735.
- [14] McKinney J D, Seo Dong-sun, Weiner A M. Direct space-to-time pulse shaping at 1.5 μm [J]. IEEE Journal of Quantum Electronics, 2003, 39(12): 1635-1644.

(上接第 194 页)

参考文献:

- [1] 中国电子商务研究中心. 农业类网站数量已超过 4 万家 [EB/OL]. 2013-04-22. <http://b2b.toocle.com/detail--6095919.html>.
- [2] 王利, 刘宗田, 王燕华, 等. 基于内容相似度的网页正文提取[J]. 计算机工程, 2010, 36(6): 102-104.
- [3] 刘峰, 王儒敬. MD5 算法在农业数据消重中的应用[J]. 计算机系统应用, 2009, 18(1): 104-106.
- [4] 郑鹏. 搜索引擎中的相似网页探测算法研究[D]. 武汉: 华中科技大学, 2008.
- [5] 李永可, 张太红, 冯向萍, 等. 中文农业网站多元线性回归识别研究[J]. 新疆农业大学学报, 2011, 34(5): 442-446.
- [6] Approximation. Lucene 中文分析器的中文分词准确性和性能比 [EB/OL]. 2009-03-06. <http://approximation.iteye.com/blog/345885>.
- [7] 何忠秀, 王霜, 安礼成. 基于向量空间的网页内容相似度计算方法研究[J]. 计算机与现代化, 2010(9): 53-55.

- [8] 李连, 朱爱红, 苏涛. 一种改进的基于向量空间文本相似度算法的研究与实现[J]. 计算机应用与软件, 2012, 29(2): 282-284.
- [9] Liu Hongzhe, Bao Hong, Xu De. Concept vector for semantic similarity and relatedness based on WordNet structure[J]. Journal of Systems and Software, 2012, 85(2): 370-381.
- [10] Landauer T K, McNamara D S, Dennis S, et al. Handbook of latent semantic analysis[M]. [s. l.]: Lawrence Erlbaum Associates, 2007.
- [11] 李瑞杰. 基于语义的网页相似性研究[D]. 郑州: 河南工业大学, 2011.
- [12] 宋涛, 施水才, 房祥, 等. 基于改进的潜在语义分析的文本聚类[J]. 北京信息科技大学学报(自然科学版), 2012, 27(3): 21-25.
- [13] 刘翔, 施干卫, 丁祖荣. 论文相似度的计算研究—基于 VSM 模型[J]. 情报杂志, 2010, 29(2): 142-144.
- [14] 肖志军, 冯广丽. 基于《知网》义原空间的文本相似度计算[J]. 科学技术与工程, 2013, 13(29): 8651-8656.

中文农业网页去重及相似度判断研究

作者：[赵涛](#)，[张太红](#)，[陈燕红](#)，[ZHAO Tao](#)，[ZHANG Tai-hong](#)，[CHEN Yan-hong](#)

作者单位：[赵涛, 陈燕红, ZHAO Tao, CHEN Yan-hong \(新疆农业大学 计算机与信息工程学院, 新疆 乌鲁木齐, 830052\)](#)，[张太红, ZHANG Tai-hong \(新疆农业大学 计算机与信息工程学院, 新疆 乌鲁木齐 830052; 中国农业大学 信息与电气工程学院, 北京 100083\)](#)

刊名：[计算机技术与发展](#)

英文刊名：[Computer Technology and Development](#)

年，卷(期)：2015(1)

引用本文格式：[赵涛](#). [张太红](#). [陈燕红](#). [ZHAO Tao](#). [ZHANG Tai-hong](#). [CHEN Yan-hong](#) [中文农业网页去重及相似度判断研究](#)[期刊论文]-[计算机技术与发展](#) 2015(1)