

中文微博的 Hashtag 话题相关性分析

胡长龙 唐晋韬 王 挺

(国防科学技术大学计算机学院 长沙 410073)

摘 要 Hashtag(微博话题词)是发布者为用户信息创建的话题标签,能帮助用户在海量微博数据中高效发现热点话题。Hashtag 由用户创建的特性使得不同的 Hashtag 可能代表着同一个话题,挖掘 Hashtag 之间的话题相关性将有助于热点话题发现和聚合展示。研究了 Hashtag 之间相关性分析问题,抽取了 Hashtag 文本特征、微博内容、Hashtag 的出现次数-时间分布以及 Hashtag 共现等一系列特征,以分析 Hashtag 之间的话题相关性。在新浪微博数据上的实验结果显示,这一系列特征组合能较好地帮助 Hashtag 相关性分析。

关键词 微博,话题相关性,Hashtag,特征抽取

Topical Relevance Analysis of Hashtags in Chinese Microblogging Environment

HU Chang-long TANG Jin-tao WANG Ting

(College of Computer, National University of Defense Technology, Changsha 410073, China)

Abstract Hashtag (the topical words of a micro-blog) is a kind of topic label of microblog created by publisher, which can help users find hot topics efficiently from the massive micro-blog data. Different Hashtags created by different publisher may describe the same topic. Thus mining the relevance between the Hashtags will help to find hot topics more efficiently. In this paper, a wide range of features were explored to analyze the topical relevance between Hashtags, such as the Hashtag text, content of the related microblog, the time of occurrence and the co-occurrences of Hashtags. The experimental results show that the proposed features are helpful for topical relevance analysis of Hashtags.

Keywords Micro-blog, Topical relevance, Hashtag, Feature extraction

1 引言

近年来,以微博为代表的社交媒体的发展非常迅速。微博是一个基于用户关系信息分享、传播以及获取的平台。用户可以通过 Web、移动设备客户端等方式便利地发布不超过 140 字的微博信息,并在社交圈中即时分享及传播。

微博等社交媒体中的信息主要由广大个人用户发布传播,因此其一方面在创造海量信息及话题,但另一方面却因信息的碎片化、冗余化而难以被有效利用。2007 年, Twitter 推出了 Hashtag(话题词)功能。在 Twitter 中, Hashtag 是以“# [a-zA-Z0-9]+”进行表示的字符串。在中文微博(新浪微博等)中, Hashtag 则是以“# 话题词 #”的形式来表示。作为微博最具特色的属性之一, Hashtag 可以帮助微博话题的聚合展示,以利于用户及研究者快速发掘热点话题。例如,用户通过点击带有超链接的 Hashtag——“# 甘肃定西地震 #”, 就可以得到一系列与 2013 年 7 月 22 日甘肃省定西市发生的 6.6 级地震相关的微博。

Hashtag 可以很好地指示微博的话题^[1], 但不同的 Hashtag 可能也会指向相同的话题。Jagan 等^[2]在对 Twitter 上的博文作在线聚类时,发现如果简单地把包含同一 Hashtag 的博文归为一类别,会导致话题的碎片化和类别的冗余。

因为对于某热点事件、新闻话题而言,由于不同发布者的描述习惯不一致或者侧重了话题的不同方面,常常有多个 Hashtag 描述相同的情况。例如,“# 甘肃定西地震救援 #”这一 Hashtag,它更多描述的是甘肃地震的救援情况。

从海量数据中挖掘热点话题一直是社会媒体中信息传播研究的一个基础问题。如果能挖掘出指向同一话题的不同 Hashtag,则可以更好地利用 Hashtag 进行微博聚类、从海量数据中提取热点话题,并将与某话题相关的 Hashtag、微博文本聚合呈现给用户,方便用户对话题的了解以及交流。

本文以新浪微博作为研究对象,分析中文微博平台上 Hashtag 之间的话题相关性。我们希望在给定若干个 Hashtag 以及相关微博的情况下,通过抽取一系列特征来判断哪些 Hashtag 描述同一个话题。很多描述同一话题的 Hashtag 会有一定的词汇重叠,或者在微博内容上有一定的相似度,但并不总是如此。如“# OZ214 航班事故 #”和“# 波音 777 坠机 #”都是描述 2013 年 6 月 7 日凌晨一架波音 777 型客机在旧金山国际机场着陆时的坠机事件。所以,本文不仅分析了较为直观的词汇特征及内容特征,还从 Hashtag 共现频率、生命周期等角度挖掘了一些统计特征以解决上述问题。实验结果显示,我们所提出的特征能较好地指示 Hashtag 的相关性。

本文第 2 节介绍了相关工作;第 3 节对研究的问题以及

本文受国家自然科学基金项目(61202337, 61170156)资助。

胡长龙(1988—),男,硕士生,主要研究方向为自然语言处理;唐晋韬(1981—),男,博士,讲师,主要研究方向为社会网络分析和自然语言处理;

王 挺(1970—),男,博士,教授,博士生导师,主要研究方向为自然语言处理。

研究思路进行了阐述;第4节给出了相关特征及计算方法的具体介绍;第5节描述了实验数据以及实验结果,并对实验结果进行了分析;最后总结了全文,并阐述了下一步工作。

2 相关工作

与本文任务非常相似的研究是新闻文本中的话题关联识别技术。话题关联识别技术来源于 TDT (Topic Detection and Tracking, 话题检测与跟踪) 任务,主要被用来判断每个报道对是否描述了同一个话题。通过建立报道和话题的表示模型、确定模型间的话题关联计算方法(即表示模型间的相似度计算)、对相似度进行整合并与决策阈值比较^[3]几个步骤计算报道对的话题关联。但由于新闻文本和微博文本在长度、语言特征、传播特征上的巨大差异,使得话题关联识别技术在微博上所取得的效果不如在新闻文本上。如果直接将包含 Hashtag 的博文当成报道,并使用该技术来判断 Hashtag 之间的相关性,有可能将一些微博内容相近似但话题不相关的 Hashtag 判断为相关。因此需要引入微博的一些属性信息来作为相关性分析的重要特征。

另一方面,已有研究者开展了 Hashtag 的聚类工作。Poschko^[4]认为如果两个 Hashtag 在一条博文上共现,那么它们可能描述了同一个话题。Dolan Antenucci^[5]在此基础上,提出了 Hashtag 共现概率的计算方法,并且设定阈值过滤部分噪音数据。Hashtag 之间有共现则表明两个 Hashtag 之间有一定程度的相关,但是没有共现并不意味着不相关。因此,我们在新浪平台数据上对该特征的有效性进行了检验,并扩展了一些其他特征来计算 Hashtag 之间的相关性。

另外,由于微博信息实时性强,在对微博进行聚类研究时,微博的发表时间常常作为一个重要因素被引入^[2,6,7],即考虑话题的爆发及消失这一过程对聚类的影响。Hashtag 相关性分析也是为了对其代表的微博文本基于话题进行聚类,其同样具有话题生命周期等时间特征,基于这个思路,我们也分析了 Hashtag 的出现及消失这一生命周期^[8]对 Hashtag 之间话题相关性识别的影响。

3 问题定义和研究思路

3.1 问题定义

本文主要研究 Hashtag 之间的相关性判别问题。我们用 Status (Time, Id, Text, User, Others) 表示一条微博,其中 Time 表示微博的发表时间,Id 是微博的 id 标识符,Text 是微博的文本内容,User 表示微博发表者的用户信息,Others 表示其他属性,如微博的转发数目、评论数目等。而本文的目标是,在给定两个 Hashtag_A 和 B 以及包含该 Hashtag 的微博集合 {Status_{A1}, Status_{A2}, ..., Status_{Am}} 和 {Status_{B1}, Status_{B2}, ..., Status_{Bn}} 的条件下,从中挖掘一组特征,学习得到它们的组合方式用以帮助判断 Hashtag A 和 Hashtag B 之间是否话题相关。

3.2 研究思路

Hashtag 之间的相关性分析实为一个二值分类问题。对于两个 Hashtag 而言,根据它们涉及的话题,可以分为相关、不相关。如果 Hashtag 之间是相关的,那么它们之间的话题内容会比较相似,相关微博发布时间的分布上也较为近似。

因此,我们从 Hashtag 本身、相关微博挖掘一组特征来进行相关性研究,这些特征包括:Hashtag 文本特征、相关的微博内容、Hashtag 出现次数-时间分布、Hashtag 的共现。通过分析这些特征的作用,学习它们的组合方式,最终判断 Hashtag 间的相关性。我们将对这些特征的效果进行评价,分析不同特征对相关性分析的影响。

本文采用基于随机森林的方法,这种模型对不平衡的数据而言,可以平衡误差^[9],适合于本文在学习特征组合时,训练数据中正反样本比例不平衡的情况。本文也对 SVM 的方法进行了研究,但效果并不如随机森林,主要原因是反例样本数目远远多于正例数目,导致对正例样本识别率降低。

4 特征设计

我们发掘了一些从微博内容及属性上计算得到的 Hashtag 话题相关性特征。这些特征以及其计算方法描述如下。

4.1 Hashtag 文本特征(JaccardOfHashtag)

在很多情况下,描述相同话题的不同 Hashtag 之间会有一些的词汇重叠如: #波音坠毁# 和 #波音 777 坠机# 两个 Hashtag 均描述波音 777 型客机坠毁事件。因此,我们首先对 Hashtag 进行了分词,然后使用 Jaccard 系数^[10]来度量 Hashtag 之间的字符重合度,其公式如下:

$$J(A, B) = \frac{|S_A \cap S_B|}{|S_A \cup S_B|}$$

式中, S_A、S_B 分别表示 Hashtag A 和 Hashtag B 分词之后得到的词集合。|S_A ∩ S_B| 表示 Hashtag A 和 Hashtag B 采用的相同词语的个数, |S_A ∪ S_B| 则表示构成 Hashtag A 和 Hashtag B 所采用词语的总个数。通过计算 Jaccard 系数,可以得到两个 Hashtag 在词上面的重合度。

4.2 包含 Hashtag 的微博内容(WeiboText)

微博的文本内容是对 Hashtag 进行的进一步描述。不同的发布者很可能采用不同的词作为话题的 Hashtag,使得描述相同话题的 Hashtag 不一定有字符上的重叠,而某些构成 Hashtag 的词在其它微博上并未用 # 标记^[11,12],如下面这条微博:

“XXXX: #OZ214 航班事故#【7月8日记者招待会内容问答内容】二.问:在旧金山机场着陆时,操作该飞机的机长仅有43小时驾驶波音777的经验。而外界认为,这是造成此次事故的间接原因。对上述内容您有什么看法?(回应“实习生”,“副机长”等的质疑)。”

这条微博使用的 Hashtag 是 #OZ214 航班事故#, 与 #波音 777 坠机# 并没有词汇重叠,但其博文中却包含了“波音 777”这样的其他 Hashtag 中的词汇。因此分析微博的文本内容可以帮助我们更好地识别 Hashtag 之间的话题相关性。

我们将包含相同 Hashtag 的微博作为一篇文档组织在一起,然后可以计算不同 Hashtag 对应的文档之间的文本相似度。本文采用 VSM (Vector Space Model, 向量空间模型) 来表示文本,特征值采用 TF-IDF 计算得到,利用余弦相似度来计算文档之间的相似度。微博内容中包含了一些影响文本相似度计算的特殊格式,例如用“@某人”的格式来表示“提到”某个用户,而这些并没有帮助文本相似度计算。所以在分词预处理阶段,除了过滤掉停用词,我们也将这些特殊格式的符

号去掉,包括表情符号、“@某人”以及 URL 短链接。

4.3 Hashtag 的出现次数-时间分布(Time)

时间作为微博的一个重要属性,表明了用户在哪天参与了某话题讨论,同时也侧面反映了 Hashtag 对涉及的话题的发生发展时间。如图 1 所示,我们列举了 2013 年 7 月 6 日至 7 月 26 日 #韩亚航空波音 777 坠机#、#波音 777 坠机#、#甘肃定西地震# 3 个 Hashtag 的出现次数,其中 #韩亚航空波音 777 坠机# 和 #波音 777 坠机# 描述 2013 年 7 月 7 日凌晨一架波音 777 型客机坠机事件, #甘肃定西地震# 则描述 2013 年 7 月 22 日在甘肃省定西市发生的地震。可以看出相同话题的 Hashtag 在同一个时间段内,其出现次数的分布比较一致,而不同话题的 Hashtag 之间,出现次数的差异也比较明显。

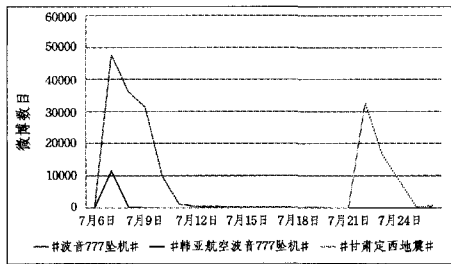


图 1 通过新浪微博搜索页面搜索得到的,2013 年 7 月 6 日至 7 月 26 日分别包含 #甘肃定西地震#、#韩亚航空波音 777 坠机#、#波音 777 坠机# 3 个 Hashtag 的微博数目

因此,我们以 $(n_{A1}, n_{A2}, \dots, n_{Ai}, \dots, n_{Am})$ 表示在 m 天内, Hashtag A 每天出现的数目。不同 Hashtag 之间时间上的相似度计算则采用余弦公式求得,表示如下:

$$S_T(A, B) = \frac{\sum_{i=1}^m (n_{Ai} * n_{Bi})}{\sqrt{\sum_{i=1}^m n_{Ai}^2} * \sqrt{\sum_{i=1}^m n_{Bi}^2}}$$

4.4 Hashtag 的共现(Co-Occurrence)

当不同的 Hashtag 同时出现在一条微博中时,这种情况称为 Hashtag 之间的共现。越是经常发生共现的 Hashtag,它们所代表的话题应该越相似。我们认为:

- 1) 共现概率越高的 Hashtag, 话题越有可能是相关的。
- 2) 如果 Hashtag A 和 Hashtag B 相关联, 那么与 Hashtag A 共现的 Hashtag 越有可能与 Hashtag B 共现。

因此,我们使用以下公式来计算 Hashtag 之间的共现概率。

$$S_{C1}(A, B) = \frac{(\frac{n_{AB}}{\sum n_{A*}} + \frac{n_{BA}}{\sum n_{B*}})}{2} \quad (1)$$

$$S_{C2}(A, B) = \frac{(\frac{m_A}{\sum n_{A*}} + \frac{m_B}{\sum n_{B*}})}{2} \quad (2)$$

式中, n_{AB} 表示 Hashtag A 和 Hashtag B 共现的次数, $\sum n_{A*}$ 表示的是和 Hashtag A 有共现关系的所有 Hashtag 的数目。另外, 对于与 Hashtag A 和 Hashtag B 两者都有共现关系的一类 Hashtag, 其与 Hashtag A 共现的总次数则表示为 m_A , 与 Hashtag B 共现的总次数则表示为 m_B 。

我们发现有一些不相关联的 Hashtag 也可能出现在同一条微博中。由于 $\sum n_{A*}$ 和 $\sum n_{B*}$ 的大小跟微博的数目直接相

关, 当微博数目少时, 式(1)和式(2)的计算很容易受这类噪音数据^[5,13,14]影响, 因此这里引入式(3), 实验过程中通过分类器来选择最佳的 Hashtag 相关性判别函数。

$$S_{C3}(A, B) = \min(\sum n_{A*}, \sum n_{B*}) \quad (3)$$

5 实验

5.1 实验数据及设置

由于对中文 Hashtag 的话题相关性研究尚无公认的语料集和标注结果, 本文使用新浪提供的 API 抓取收集了从 2013 年 3 月到 4 月共两个月时间的微博数据。考虑到热门话题相关的 Hashtag 会比较多, 为了增加正例的数量, 我们从中选取了 130 个与热门话题相关的 Hashtag, 相关涉及的微博数目共 55533 条(含同一个 Hashtag 的微博数目最少为 30 条), 生成 $C_{130}^2 = 8385$ 个 Hashtag 对。由于 Hashtag 之间的相关性判断类似于 TDT 任务中的话题关联识别, 因此我们参照 TDT4 标注大纲对数据进行标注。如果两个 Hashtag 讨论相同话题, 则认为相关, 如果两个 Hashtag 讨论完全不同的话题, 则认为无关, 如果存在疑虑, 则往往认为相关。我们借助 3 位志愿者对这些 Hashtag 对进行人工标注。首先由两位志愿者进行独立标注, 当标注的结果存在差异时, 则以第三位志愿者的判断作为标注结果。最终, 我们得到正例 547 对, 反例 7838 对。考虑到正反例样本不平衡的问题, 我们利用随机森林训练相关性识别函数, 以平衡误差。为了避免过度拟合, 我们对数据集的实验采取了 5 折交叉验证。实验结果采用准确率、召回率、F-值来度量。

为了考察各个特征的作用, 我们逐步将特征累加, 形成了如下几个测试方法:

方法 1(baseline) 仅使用 Hashtag 本身的字符重合度(JaccardOfHashtag)这一特征。

方法 2 在方法 1 基础上添加包含 Hashtag 的微博内容(WeiboText)这一特征。

方法 3 在方法 2 基础上添加 Hashtag 的出现次数-时间分布(Time)这一特征。

方法 4 在方法 3 基础上添加 Hashtag 的共现(Co-Occurrence)特征, 即包含了上面列举的所有特征。

5.2 实验结果

表 1 显示了上述测试方法的实验结果。实验结果显示, baseline 方法有较高的准确率, 但是召回率较低。这是因为该方法不能发现在词汇上没有重叠的 Hashtag 之间的话题相关性。方法 2 由于引入了微博内容特征, 借助微博内容的相似性判断不同 Hashtag 的话题相关性, 提高了召回率。但有些相似话题的微博内容近似, 如“#台湾地震#”和“#雅安平安#”, 这会导致不相关联的 Hashtag 被判断为相关, 从而降低了准确率。方法 3 中引入了时间特征, 可以区别不同时间段的不同话题, 并可以提高同一时间段发生的话题的关联度, 从而在准确率和召回率上均有提高, 如该特征能很好地区分发生在 2013 年 3 月份的“#台湾地震#”和发生在 2013 年 4 月底的“#雅安平安#”话题。方法 4 在方法 3 基础上, 加入了 Hashtag 的共现特征。实验结果表明, 在中文数据上, Hashtag 的共现和英文数据一样是判断 Hashtag 话题相关性的一

(下转第 245 页)

源的格式,如 PDF 等,实现内容的有效传递和知识共享。在 IR 设计中尽量采用列表框选择数据的方法,如语言、学科、主题词、资源类型等规范化项目供资源提供者选择,尽量减少需要其录入的条目^[11],还可安排专人协助提交存储,辅助转换格式、纠正拼写错误、完成提交程序等。

此外,应重视 IR 服务功能的扩展,开发基于浏览、标记、订阅、检索、评论等满足个性化服务需求的增值功能,包括创建个人出版物列表、统计论文点击率和引用率等,如为作者提供建立和维护个人出版物目录的服务。还应着重考虑在 IR 中嵌入 RSS 订阅服务,为特定社区、特定作者及特定主题等提供多个站点实时更新的 RSS 源文件,并利用信息追踪技术和检索记忆理念,使科研人员输入的检索词为系统所记忆和匹配,进而实现相关资源的主动推送。

结束语 高校应通过多方渠道提高 IR 的知名度和认同感,使实施对象突破图书情报领域的实践者。积极借鉴国外 IR 建设经验,充分发挥图书馆辅助协调管理的主体角色,更加注重版权许可、质量控制、个性化增值服务及标准化建设等内容,促进学术交流和资源共享,扩大高校争取人才和资金的优势,提高核心竞争力。

参考文献

[1] “开放获取 (OA)”推动信息共享 [EB/OL]. <http://www.peo->

- ple.com.cn/GB/Paper464/15160/1344583.html,2013-08-08
- [2] Budapest Open Access Initiative [EB/OL]. <http://www.soros.org/openaccess/read.shtml>,2013-08-11
- [3] 孙振良. 高校机构知识库建设现状及策略研究[J]. 情报科学, 2010(03):353
- [4] Crow R. The Case for Institutional Repositories; A SPARC Position Paper. [EB/OL]. <http://www.arl.org/sparc/IR/ir.html>,2013-08-13
- [5] ROAR: Registry of Open Access Repositories [EB/OL]. <http://roar.eprints.org/>,2013-08-10
- [6] OpenDOAR:Directory of Open Access Repositories [EB/OL]. <http://www.opendoar.org/countrylist.php?cContinent=Asia#China>,2013-08-14
- [7] 厦门大学学术典藏库 [EB/OL]. <http://dspace.xmu.edu.cn/dspace/>,2013-08-16
- [8] Cybermetrics Lab. Ranking Web World Repositories[EB/OL]. <http://repositories.webometrics.info/en/Asia>,2013-08-15
- [9] 台湾大学机构典藏[EB/OL]. [2013-08-13]. <http://ntur.lib.ntu.edu.tw/>
- [10] Swan A, Brown S. Open access self-archiving: an author study [EB/OL]. <http://cogprints.org/4385/>,2013-08-13
- [11] 郎庆华. 机构知识库自存储资源的获取策略研究[J]. 情报杂志, 2009(7):169

(上接第 237 页)

个有效特征,通过加入该特征在准确率和召回率上均有提高。

表 1 不同的特征组合对 Hashtag 相关性检测的影响,符号 v 或 * 表示实验结果在显著水平 0.05 下优于(v)或不如(*)base-line 方法

	方法 1 (baseline)/%	方法 2/%	方法 3/%	方法 4/%
准确率	0.879	0.786 *	0.886	0.896
召回率	0.643	0.789v	0.868v	0.876v
F-值	0.741	0.787v	0.877v	0.886v

结束语 Hashtag 的话题相关性分析有助于从海量的微博信息中有效地挑选用户可能感兴趣的微博、帮助热点话题发现和聚合展示、方便用户沟通交流。本文根据微博文本的特点,不仅考虑了构成 Hashtag 的词汇重合度,还考虑了相关微博内容、Hashtag 的出现次数-时间分布、Hashtag 共现等特征来帮助 Hashtag 相关性分析。实验结果表明本文抽取的一系列特征都有助于 Hashtag 相关性判断。

由于存在着一些不相关的 Hashtag 共现的情况,在未来工作中,我们将考虑如何过滤这类噪音数据。另外,我们将挖掘其他特征来帮助判别 Hashtag 之间的相关性,如度量 Hashtag 之间的 google 距离^[15]等。将 Hashtag 相关性判别技术用于微博聚类、Hashtag 推荐等也是未来工作之一。

参考文献

- [1] Rosa KD, Shah R, Lin B, et al. Topical clustering of tweets[C]// Proceedings of the ACM SIGIR/SWSM, 2011
- [2] Sankaranarayanan J, Samet H, Teitler B E, et al. Twitterstand: news in tweets[C]// Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2009:42-51

- [3] 张晓艳. 新闻话题表示模型和关联追踪技术研究[D]. 长沙:国防科学技术大学, 2010
- [4] Pöschko J. Exploring Twitter Hashtags[Z]. 2011
- [5] Antenucci D, Handy G, Modi A, et al. Classification of Tweets Via Clustering of Hashtags[Z]. 2011
- [6] 郑斐然, 苗夺谦, 张志飞. 一种中文微博新闻话题检测的方法[J]. 计算机科学, 2012, 39(1):138
- [7] Cataldi M, Di Caro L, Schifanella C. Emerging topic detection on Twitter based on temporal and social terms evaluation[C]// Proceedings of the Tenth International Workshop on Multimedia Data Mining. ACM, 2010:4
- [8] Chang H C. A new perspective on twitter Hashtag use: diffusion of innovation theory[J]. Proceedings of the American Society for Information Science and Technology, 2010, 47(1):1-4
- [9] 随机森林-维基百科, 自由的百科全书[DB/OL]. <http://zh.wikipedia.org/wiki/随机森林>, 2013
- [10] Leydesdorff L. On the normalization and visualization of author cocitation data; Salton's Cosine versus the Jaccard index[J]. Journal of the American Society for Information Science and Technology, 2008, 59(1):77-85
- [11] Laniado D, Mika P. Making sense of twitter[M]. The Semantic Web-ISWC 2010. Springer Berlin Heidelberg, 2010:470-485
- [12] Guo W, Li H, Ji H, et al. Linking Tweets to News: A Framework to Enrich Short Text Data in Social Media
- [13] Wang A H. Don't follow me; Spam detection in twitter[C]// Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on. IEEE, 2010:1-10
- [14] Benevenuto F, Magno G, Rodrigues T, et al. Detecting spammers on twitter[C]// Collaboration, electronic messaging, anti-abuse and spam conference (CEAS). 2010
- [15] Cilibrasi R L, Vitanyi P M B. The google similarity distance[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(3):370-383