博客中重复评论发现*

刁宇峰，王 昊，林鸿飞，杨 亮
大连理工大学 信息检索研究室，大连 116024
E-mail: sin19871028@163.com

摘 要：随着近年来互联网的迅猛发展，Blog 上的数据呈现爆炸式的增长，产生了大量的重复评论，这些重复评论对观点挖掘、信息跟踪、搜索引擎等 Web 应用的处理带来了严峻的问题。本文针对 Blog 中评论本身的特点，提出一种有效的结合主题信息的 TopicSig 算法去检测 Blog 中的重复评论。该方法主要针对博客中的所有评论进行主题抽取，并结合高频词共同作为特征先行词，以抽取改进的 Shingle 特征，高度概括评论的核心内容，使用有效的相似度算法比较，从而发现博客中的重复评论。经实验显示，该方法可以发现大多数重复评论，实验结果显示取得了较好的结果，使 Blog 信息更加准确、有效的为用户使用。

关键词：Blog；重复评论；主题；TopicSig；Shingle

Blog Opinion Near-duplicate Discovering

Diao Yufeng,Wang Hao, Lin Hongfei, Yang Liang
Information Retrieval Laboratory, Dalian University of Technology, Dalian 116024
E-mail: sin19871028@163.com

Abstract: The data of the Blog show explosive growth, which results from a large number of near-duplicate comments, with the recent rapid development of Internet. It poses a serious problem on the views of opinion mining, information tracking, search engine and other Web application processing. In this paper,we propose an effective feature extraction algorithm—TopicSig algorithm,which combined with the topic to discover the near-duplicate comments aiming at the characteristics of Blog.This method extracts the topics as the feature antecedent, then extracts the improved Shingle feature to highly summary the core content of the comments .Shown by the experiment, this method can find the most near-duplicate opinions, and the experiments show good results, it can make the Blog information more effectiveness for users.

Keywords: blog; near-duplicate opinion; topic; TopicSig; Shingle

1 引言

在现今，重复信息的研究已经成为一个重要的研究领域[1,2,3,4]。这些重复信息的存在有以下几个弊端：从搜索结果质量来看，搜索引擎提供重复结果严重伤害了用户体验；从收录页面质量来看，重复转载更容易引入死链接和垃圾信息；从资源占用和维护角度来看，存储重复信息即浪费硬盘又不利于网页的更新；另外重复信息转载是一定程度上对原创用户的不公平。现有的研究工作主要在于如何进行文档级别的重复拷贝检测并取得了不错的成果。但是目前仍然存在一些问题，典型的例子为文档中抄袭检测、引文部分的拷贝检测、网页中重复内容检测问题等拷贝检测，这类问题是段落级别和句子级别的重复检测，因而也无法使用文档级别的拷贝检测方法有效的检测出来。文献[5,6,7,8,9,10]等都可以部分的解决此任务，但是经研究发现，仅依靠相似度和文档种类这类计算相似度的方法不足以为所有应用提供充分的信息。本文分析重复评论的特点，在新浪博客评论语料集上，对所有的评论使用 LDA 模型挖掘出隐含的主题信息，将这些主题信息结合常见词、

* 基金项目：国家自然科学基金资助项目（编号：60673039，60973068）、国家 863 高科技计划资助项目（编号：2006AA01Z151）、教育部留学回国人员科研启动基金和高等学校博士学科点专项科研基金资助课题（编号：20090041110002）。

情感倾向性词等一起作为特征先行词，以抽取改进的 Shingle[11]特征，这样可以高效的覆盖评论的核心内容，在对两个评论的特征集合进行相似度计算时，采用最短编辑距离和 HowNet 结合的方法，进而发现 Blog 中的重复评论，便于用户的阅读和使用，也为基于 Web 信息的应用打下了良好的基础。

本文具体方法在下面详细介绍：第二部分主要介绍相关工作；第三部分主要介绍核心算法—基于 LDA 的重复评论发现，并使用最短编辑距离和 HowNet 结合的方法比较相似度；第四部分是实验流程以及结果分析；最后，在第五部分中总结工作并计划下一步工作。

2 相关工作

Latent Dirichlet Allocation(LDA)模型是 Blei 等在 2003 年提出的[13]，属于主题模型(Topic Models，是当前文本表示研究的主要范式)的一种。作为一种产生式模型，LDA 模型已经成功的应用到文本分类，信息检索等诸多文本相关的领域[14][15]。

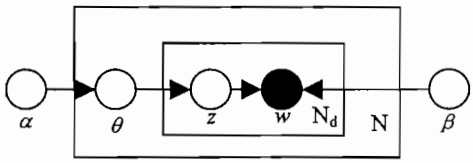

图 1 LDA 的图模型表示形式

LDA 是一个多层的产生式全概率生成模型，是典型的有向概率图模型，如图 1 所示，包含词，主题和文档三层结构。LDA 模型的生成过程是一种概率抽样过程，是一种基于潜在主题生成文档中词的过程，其生成过程如下(K 为主题个数，N 为文档个数)：

1、从 Dirichlet 先验 β 中为每个主题抽取多项式分布 φ，共抽取 K 个分布。

2、从 Dirichlet 先验 α 中为每个文档抽取多项式分布 θ，共抽取 N 个分布。

3、对语料库中的所有文档 N 和文档中的所有词 W：

1) 从多项式变量 φ 中抽取主题 z；

2) 从多项式变量 θ 中抽取词 w。

Border 在 1997 年提出一种 DSC[11](Digital Syntactic Clustering)算法。该算法将文本按照 n 字一组组合起来称为一个 Shingle，而整个文本则由 Shingles 集合组成。比较使用 Jaccard 算法来衡量两个集合间的相似度。这种筛选方法虽然简单，但却极大地影响到了算法的精确性。2008 年，M.Theobald[7]等人提出的 SpotSig 算法在文档级别的拷贝检测中取得了很好的效果。该算法以停止词作为先行词，比较时亦使用 Jaccard 相似度，并提出了基于文档特征向量长度的剪枝方法，大大提高了 SpotSig 在比较阶段的效率。Zhang[9][10]等人在 2010 年提出一种改进型的句子级别的文本特征提取方法 Low-IDF-Sig 算法，该算法选取一些具有最低逆向文档频率的词汇作为特征先行词。在比较时，使用基于倒排索引进行剪枝的方法，缩减了比较时间，实验结果证明该算法是行之有效的。

3 TopicSig 算法

本文在 Blog 领域内进行研究，主要考虑的是评论的文本信息。针对评论的特点进行分析，提出一种改进算法 TopicSig 算法，本算法采用主题模型抽取评论集合的主题信息，将主题信息、情感词和常见词按一定比例结合构成的词汇作为特征先行词，用以代表整条评论。评论间比较时，不采用 Jaccard 算法，而是采用最短编辑距离和 HowNet[17]结合的方法。

3.1 特征先行词选取

经分析 Blog 评论的特点，发现评论中抽取的特征先行词主要有三类，具体见表 1 所示(外部资源是大连理工大学信息检索实验室的情感词汇本体[12]，简称情感本体)：

表 1 特征先行词集合

先行词类别	主要元素
情感倾向性词	情感词典、程度副词、否定词
常见词	最低逆向文档频率的常见词汇
主题词	评论隐含主题信息

接下来的关键就在于如何按比例选取三类特征先行词进行融合作为最终的特征先行词。这里，主要采用两种方法进行融合。

（1）基于线性的特征先行词融合

经分析 Blog 评论，发现本文提出的情感倾向性词、具有最低逆向频率的常见词和主题词大部分出现于评论，广泛的被用户在评论中使用，则最终的先行词集合组成如下：

$$A=\{Sen\_Set + IDF\_Set + Topic\_Set\} \quad (1)$$

其中，A 为最终的先行词集合，Sen_Set 为情感强度高于阈值 λ_1 的情感词集合，IDF_Set 为 IDF 值低于阈值 λ_2 的常见词集合，Topic_Set 为主题排名在前 n 的且在该主题下排名前 m 的主题词集合，其中 λ_1 和 λ_2 均在(0,1)之间，n、m 为整数。

（2）基于主题检索模型的特征先行词融合

由于 Blog 评论特点，本文受到文[18]中的方法启发，采用概率检索模型来发现重复评论，该方法不用着重筛选特征集合进行训练。将评论和候选特征先行词的问题看作是检索问题，候选特征先行词为表 1 共同构成。候选特征先行词 A 假设为查询串，评论 C 当作文档，评论集合看作文档集合，在未引入主题信息前建立简单的概率检索模型 [18]，公式如下：

$$P(A|C)=\prod_{w\in A} P(w|C) \quad (2)$$

其中，C 为评论集合，A 是候选特征先行词，w 是 A 中的一个词，假定 C 中词与词之间相互独立，P(A|C)为 C 产生 A 的概率，P(w|C)为 w 在 C 中出现的概率。

在上述模型中未考虑到评论本身的稀疏性以及词之间隐含的主题信息，也未对 P(w|C)这项进行平滑，有待改进。本文在上述概率检索模型的基础上，加入引入主题模型 LDA 后发现的隐含主题集合[18]，用于进行平滑 P(w|C)，即主题检索模型。本文结合隐含的主题信息，共同建立主题检索模型。具体公式如下：

$$P(A|C)=\lambda\prod_{w\in A} p(w|C)+(1-\lambda)\prod_{w\in A}\sum_{t\in t_B}\prod_{w\in t} p(w|t)*p(t|C) \quad (3)$$

其中，t_B 为评论集合 C 的主题集合，t 为 t_B 中的一个主题，λ 为参数，p(w|t)为词 w 在主题 t 中出现的概率，p(t|C)为主题 t 在评论集合 C 中出现的概率，P(w|C)为 w 在 C 中出现的概率。

$$P(w|C)=(n_w+1)/N \quad (4)$$

其中 n_w 为词 w 在评论集合 C 中出现的次数，N 为评论集合 C 中总词数。对 Blog 中所有评论均计算 P(A|C)建立主题检索模型，若此概率大于某阈值则判定为特征先行词，反之不是。

3.2 特征标记

本算法的特征 s 的标记方法同 SpotSig 算法和 Low-IDF-Sig 算法，表示为一条紧跟在一个特征先行词 a 后的长度为定值 b 的词链，此词链的间隔为固定值 c，即使用标记 a(b,c)来表示一个特征先行词为 a，词链长度为 b，取词间隔为 c 的 TopicSig 特征 s。其具体差别如下：

（1）TopicSig 特征在选取特征先行词时，根据 3.1 得到的外部资源特征先行词表作为 TopicSig 特征的先行词，并且为确保每条评论至少有一个特征，本文将选取评论的第一个词作为特殊的先行词。

（2）TopicSig 特征构成 Shingle 时，词链既包括先行词后提取的词，也包含其本身。经研究发现，使用不同的介词、冠词、助词和叹词，评论仍表示相同的意思，因此在构成词链时，介词、冠词、助词和叹词等这类词不会在同一条词链中出现。

3.3 相似度比较

给定一个评论集合，重复拷贝检测需要进行评论间“两两”的相似度。大多数算法均使用常见的 Jaccard 算法进行比较。但是在本文，经分析 Blog 评论，该算法的精度有待提高，因此，本文提出一种最短编辑距离和 HowNet[17]结合的方法计算相似度。在计算相似度时使用检索领域中的倒排索引结构 [7]。使用倒排索引进行评论比较时，至少含有一个共同的特征的评论才会被拿出来计算相似度，这样就可以大大的减少比较的次数，提高比较的效率。

4 实验结果及分析

实验的语料来自新浪博客下载的评论，作者为博客总人气排行榜前 10 名，选取其部分博文共 100 篇博文(每人 10 篇)并从中选取评论共有 8980 条，经标注，共发现重复评论 1074 条。

图 2、图 3 显示了 3-Shingles，4-Shingles，5-Shingles 在使用 Jaccard 算法和本文提到的最短编辑距离结合 HowNet 两种相似度方法，在调整相似度阈值 τ 时 F1 值的变化趋势，可以看出使用第二种的相似度算法的 F1 有提高，并分别在相似度阈值 τ =0.6 时获得最高 F1 值 0.91 和 0.92，但是 3-Shingles，4-Shingles，5-Shingles 这三类切分方法在评论这类句子级检测上表现尤为接近，没有明显的区分且时间消耗过大，不太适合处理实时性需求强的大规模语料。

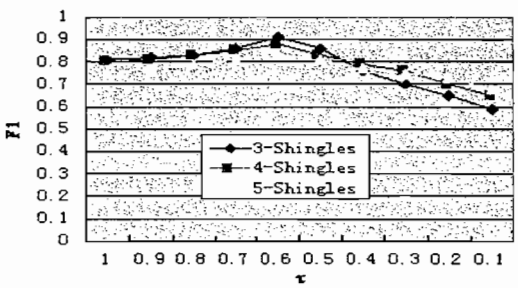

图 2　Shingle_Jaccard 算法

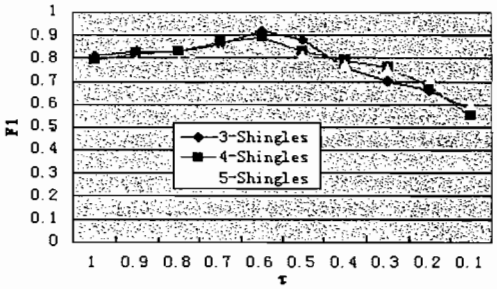

图 3　Shingle_Hownet_Edit 算法

图 4、图 5 显示了 SpotSig 算法在不同特征先行词下(停用词、停用词和情感倾向性词线性融合、停用词和情感倾向性词以及主题词线性融合、三者主题检索模型融合四类)在上述两种相似度比较下，随 τ 的变化下 F1 的变化，其中，选取词链长度为 3，词间隔为 1。M.Theobald 等人发现先行词只需选取 24 个停止词就可以取得很好的 F1 值。但是本文发现仅 24 个先行词不能从评论中抽取充分的特征，最终选取 754 个词(包括停用词、情感倾向性词和主题词)作为先行词，采用后种相似度比较，此时的 F1 达到最高点，并在阈值 τ =0.65 时 F1 最高为 0.78。

图 6、图 7 显示了 Low-IDF-Sig 算法在不同特征先行词下(常见词、常见词和情感倾向性词线性融合、常见词和情感倾向性词以及主题词线性融合、三者主题检索模型融合四类)，上述两种相似度比较的情况下，在相似度阈值 τ 的变化下 F1 的变化，其中，词链长度为 2，词间隔为 1。Zhang 等选取 500 个先行词，阈值 τ =0.6 时取得最好的 F1 值。但是我们发现仅 500 个先行词不足以抽取充分的特征，因此 F1 值还有待提高，而本文最终选取的 850 个词(其中包括停用词、情感倾向性

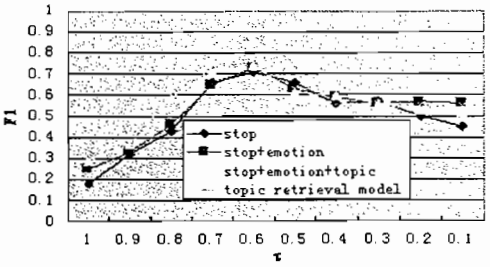

图4 SpotSig_Jaccard 算法

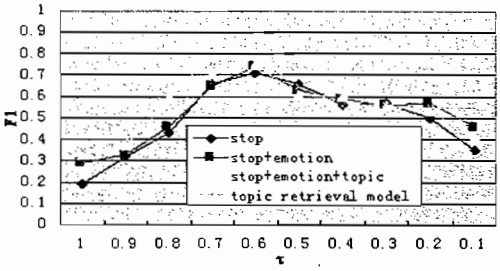

图5 SpotSig_Hownet_Edit 算法

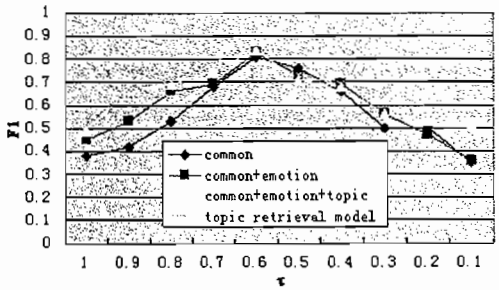

图6 Low-IDF-Sig_Jaccard 算法

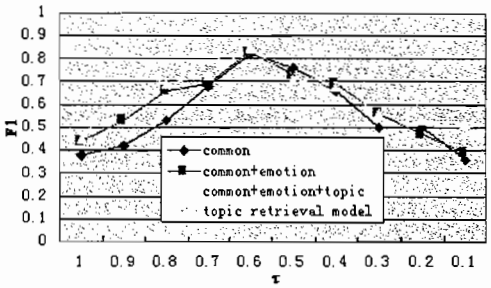

图7 Low-IDF-Sig_Hownet_Edit 算法

词和主题词)作为先行词，采用最短编辑距离和 HowNet 结合的方法计算相似度，此时的 F1 达到最高点，并在阈值 τ =0.6 时 F1 最高为 0.87。

图 8、图 9 显示了 TopicSig 算法在不同特征先行词下(常见词、常见词和情感倾向性词线性融合、常见词和情感倾向性词以及主题词线性融合、三者主题检索模型融合四类)，上述提到的两种相似度比较的情况下，在相似度阈值 τ 的变化下 F1 的变化趋势，其中，词链长度为 2，词间隔为 1。从两图可以发现，在阈值 h 的变化中，本文的相似度方法总比 Jaccard 算法效果好。在图 8 中，当阈值 τ = 0.6 时达到最高的 F1 值为 0.91，此时选取的先行词为 800 个。对于先行词为 754 个时，F1 值在 τ = 0.55 时达到最大值 0.88；而先行词取 850 个时，F1 值的最大值出现在 τ = 0.6 时，为 0.87。并且可以看出当特征先行词不同时，F1 值随 τ 的变化趋势基本相同。对于上述八个图表，F1 值的最高点都是使用的特征先行词均基于主题检索模型对三类词(常见词、情感倾向性词和主题词)进行融合得到的，相似度比较均使用最短编辑距离和 HowNet 结合的方法。无论是两种相似度比较，还是不同特征先行词集合，亦是 4 种算法，F1 值随相似度阈值 τ 的变化趋势基本相同，并且 TopicSig 算法的 F1 值最高(此时采用基于主题检索模型的先行词融合方法，最短编辑距离和 HowNet 结合的相似度算法)，证明这种算法解决 Blog 重复评论发现这一问题是有效的。

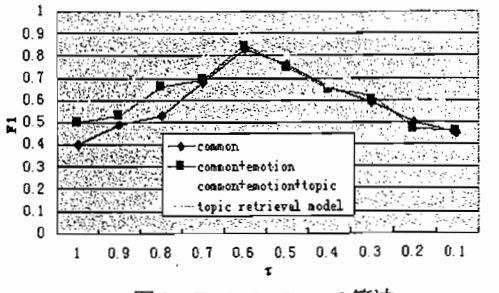

图8 TopicSig_Jaccard 算法

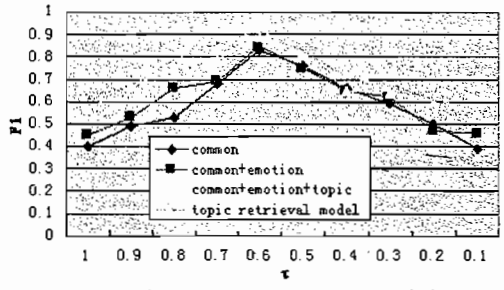

图9 TopicSig_Hownet_Edit 算法

最后，我们给出所有算法的各指标进行比较，具体见表 3：

从表 3 的各特征在新浪博客评论语料集上可以看出，使用本文提出的最短编辑距离和 HowNet

表3　各方法对比

特征算法	τ	先行词数目	F1_Jaccard	F1_EH
3-Shingles	0.6	无	0.91	0.92
SpotSig	0.65	754	0.78	0.76
Low-IDF-Sig	0.6	850	0.86	0.87
TopicSig	0.55	754	0.85	0.88
TopicSig	0.6	800	0.89	0.91
TopicSig	0.6	850	0.84	0.87

结合的方法计算相似度，在各算法上的F1值均高于Jaccard算法。3-Shingles算法的F1值在两种相似度比较下均取得了最高值，但是对于TopicSig算法来说，优势不明显，且时间消耗、空间消耗过大。而TopicSig算法的F1值明显高于SpotSig算法，证明SpotSig算法抽取的特征未能有效的表现出评论的核心内容，证明该算法更适用于句子级的特征抽取任务。TopicSig算法在F1值上高于Low-IDF-Sig算法，尤其是在特征先行词等情况均相同时F1值亦高于Low-IDF-Sig算法，证明TopicSig算法针对Blog评论考虑到的细节是有必要的相比更适用句子级的拷贝检测任务。最后从表3中可以看出，TopicSig算法在先行词从754增长到850时，F1值在先行词为800个时获得最大值，先行词数目从754个到800个时，F1值只是略有上升达到最高点，但相对空间和时间上占有确明显上升。综上所述，TopicSig算法可以很好的抽取句子级别文本的特征，表示其核心内容，适用于重复检测拷贝任务。

5　结束语与下一步工作

本文的主要工作在于识别句子级别的博客评论是否重复。我们提出TopicSig算法解决该问题。首先对博客中的所有评论使用LDA模型挖掘出其隐含的主题信息，将这些主题信息结合常见词、情感倾向性词等作为特征先行词，并抽取构成改进的Shingle特征，以高效的覆盖评论的核心内容，最后对评论特征集合进行相似度比较，这里我们采用结合最短编辑距离和HowNet的方法进行相似度比较，进而发现Blog中的重复评论。在Blog这个开放性平台，评论者可以自由发表言论，很多评论都是由诗歌、散文等隐式表达情感构成的，需要通过更深层次的方法来挖掘其隐式含义。这也是本文下一步需要解决的工作。目前，博客中的评论在研究方面的语料还不够丰富，因此，本文的语料是作者手工收集和整理，语料的丰富和校验工作还需进一步进行。以上情况都有待作进一步细致的研究。

参 考 文 献

[1] K. Muthmann, W. M. Barczynski, F. Brauer, and A. Loser. Near-duplicate detection for web-forums. In IDEAS '09, pages 142-151, New York, NY, USA, 2009. ACM.

[1] N. Shivakumar and H. Garcia-Molina. Scam: A copy detection mechanism for digital documents. In Digitial Library, 1995.

[3] N. Shivakumar and H. Garcia-Molina. Finding near-replicas of documents and servers on the web. In Proceedings of WebDB 1998, pages 204-212, London, UK, 1999. Springer-Verlag.

[4] J. Lin. Brute force and indexed approaches to pairwise document similarity comparisons with mapreduce. In Proceedings of SIGIR'09, pages 155-162, New York, NY, USA, 2009. ACM.

[5] M. Bendersky and W. B. Croft. Finding text reuse on the web. In WSDM'09, pages 262-271, New York,NY, USA, 2009. ACM.

[6] J. Seo and W. B. Croft. Local text reuse detection. In SIGIR'08, pages 571-578, New York, NY, USA, 2008. ACM.

[7] M. Theobald, J. Siddharth, and A. Paepcke. Spotsigs:robust and efficient near duplicate detection in large web collections. In SIGIR'08, pages 563-570, New York, NY, USA, 2008. ACM.

[8] A. Kołcz, A. Chowdhury, Lexicon randomization for near-duplicate detection with I-Match, The Journal of Supercomputing, v.45 n.3, p.255-276, September 2008.

[9] Qi Zhang, Yue Zhang, Haomin Yu.Efficient Partial-Duplicate Detection Based on Sequence Matching.In SIGIR'10, pages 675-682, Geneva, Switzerland, 2010. ACM.

[10] 俞昊旻，张玥，张奇，黄萱菁.基于 Low-IDF-SIG 的句子重复检测.CCIR2010,pages 24-31, 牡丹江, 中国, 2010.

[11] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig. Syntactic clustering of the Web. Computer Networks, 29(8-13): pages 1157-1166, 1997.

[12] 徐琳宏, 林鸿飞, 潘宇, 等. 情感词汇本体的构造[J]. 情报学报, 2008, 27(2): 180-185.

[13] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3: pages 993-1022, January 2003.

[14] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, ChengXiang Zhai.Topic Sentiment Mixture:Modeling Facets and Opinions In WWW 2007, pages 171-180Banff, Alberta, Canada, 2007.

[15] Yue Lu, Chengxiang Zhai.Opinion Integration Through Semi-supervised Topic Modeling.In WWW2008, pages 121-130, Beijing, China, 2008.

[16] A.Z.Broder. Identifying and filtering near-duplicate documents. In Proceedings of COM2000, page1-10, London, UK, 2000.

[17] 刘群, 李素建. 基于《知网》的词汇语义相似度计算. http://www.keenage.com

[18] X. Wei, and W. B. Croft.LDA-based document models for ad-hoc retrieval.In SIGIR'06, pages 178-185,Seattle, USA, 2006.