信息检索中的聚类分析技术

刘远超 王晓龙 刘秉权 钟彬彬
(哈尔滨工业大学 计算机科学与技术学院 哈尔滨 150001)

摘 要 信息检索/搜索引擎技术的快速发展使得信息的查全率有较大提高，而查准率以及人们获取信息的效率改善却不明显。文本聚类和多文档关键词的自动生成技术将有助于解决这一问题。其基本思想是对检索到的部分文档进行聚类处理，并对每类文档自动生成关键词，从而帮助用户判断各个类别的文档和检索需求是否相关。该文提出文档相关度和类别相关度的概念，并利用词频信息以及知网(HOWNET)中词的概念计算模型计算类别相关度，将其作为聚类合并的依据。信息获取的仿真实验表明文档检索效率有较大提高。

关键词 文档聚类, 关键词抽取, 知网, 文档相关度
中图分类号：TP391 **文献标识码**：A **文章编号**：1009-5896(2006)04-0606-04

The Clustering Analysis Technology for Information Retrieval

Liu Yuan-chao Wang Xiao-long Liu Bing-quan Zhong Bin-bin
(*School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001, *China*)

Abstract The rapid development of Information Retrieval(IR) and search engine improves recall rate greatly, whereas the enhancement on both precision rate and information retrieval efficiency is not clear. The research on document clustering and multi-document keyword extraction will help solve this problem. The basic idea is to cluster part of the documents returned by search engine, and automatically extract some keywords for each cluster. Thus user can judge whether the documents in each cluster are relevant to his need. In this paper the concept of document relevancy and cluster relevancy are proposed, and both word frequency and the concept relevancy model of HOWNET are used to compute cluster relevancy, which is used to guide the merging process of clusters. The experimental results show that the IR efficiency has improved greatly.

Key words Document clustering, Key words extraction, HOWNET, Document relevancy

1 引言

随着网络上文本信息的爆炸式增长，如何提高信息获取的效率成为研究人员广泛关注的重要课题[1, 2]。目前搜索引擎/信息检索等技术的研究已进入实用化阶段，对于用户输入的查询词，搜索引擎一般会在几秒钟内返回相应的结果。然而用户要找到自己真正感兴趣的信息仍然很难。通过对搜索引擎返回的结果进行文本聚类/分类等后处理，使结果变得更加有条理，将有助于提高信息检索的效率。多篇文档的关键词可以对同类别的文档提供介绍和描述，用户通过浏览这些信息，就可以对同一类别的所有文档有一个大致的了解，从而有效地缩小查找的范围。

本文主要研究和探讨如何对搜索引擎返回的部分结果进行有效地聚类，并自动生成各个类别的关键词。为了克服传统文档聚类方法中由于单纯采用词频信息计算文档相似度导致语义信息不够充分的问题，本文提出文档相关度和类别相关度的概念，并将类别相关度作为聚类中类别合并的依据。对同类文档自动抽取关键词的目的是生成类别描述，从而帮助用户判断类别中的文档和检索需求是否相关。

2005-01-10 收到，2005-09-26 改回
国家自然科学基金重点项目(60435020)资助课题

2 文档聚类和多文档关键词自动生成技术

传统的搜索引擎一般将返回结果线性排列。由于返回的结果中有相当一部分不是用户所需要的，并且用户难以判断文档的相关性，不得不依次打开察看，因此加重了用户的负担。通过对返回的部分重要信息(如排在前面的部分文档)进行聚类处理和关键词分析，则可以使返回结果变得更加有条理，使用户迅速找到所需的信息。而通过类别关键词的自动生成，还可以对原始查询进行扩充和精确化。

为了便于阐述，将本文中涉及到的一些术语解释如下：

定义 1 词相关度 $\mathrm{Rel}_w(w_1,w_2)$ 表示任何两个词w_1和w_2是否领域相关，其取值为 0 或 1。取值为 1 表示两个词之间具有领域相关性，取值为“0”表示不具有领域相关性。如“医生”和“患者”二词的相关度为 1，而“医生”和“棉衣”二词之间的相关度为 0。

定义 2 文档相关度 $\mathrm{Rel}_d(d_1,d_2)$ 表示文档d_1和d_2之间的领域相关程度，且 $0 \le \mathrm{Rel}_d(d_1,d_2) \le 1$。文档相关度的具体计算方法如式(3)所示。

定义 3 类别相关度 $\mathrm{Rel}_c(C_1,C_2) = \sum_{d_1 \in C_1, d_2 \in C_2} \mathrm{Rel}_d(d_1,d_2) / (|C_1| \cdot |C_2|)$ 类别间相关度的定义采用UPGMA[3]方法。这

里d_1, d_2分别表示类C_1, C_2中的文本，$|C|$表示类别C中文档的个数。

定义 4　**文档相似度** $\mathrm{Sim}_d(d_1,d_2)$　表示两篇文档d_1和d_2在向量空间模型下的文档向量夹角的余弦值，且$0 \le \mathrm{Sim}_d(d_1,d_2) \le 1$，其计算方法如式(2)所示。一般说来，两篇文档的公共词越少，其相似度越小。

定义 5　**类别相似度** $\mathrm{Sim}_c(C_1,C_2) = \sum_{d_1 \in C_1, d_2 \in C_2} \mathrm{sim}_d(d_1,d_2) / (|C_1| \cdot |C_2|)$　这里d_1, d_2分别表示类C_1, C_2中的文本，$|C|$表示类别C中文档的个数。

不难看出，类别之间的关系运算(相似度或者相关度)依赖于文档之间的关系运算，而文档之间的关系运算依赖于词之间的关系运算。本文中涉及的词相关度的计算方法来自于知网[4]中基于词的概念相关计算模型。

2.1 文档聚类

通过聚类处理，可以将同类文档组织在一起。本文采用性能较好的AHC层次聚合聚类算法[5]进行文档聚类。一般说来，在AHC算法的聚合过程中总有一次的聚类结果比较符合真实的类别划分，因此需要找到最佳的类别划分。而通过计算类别划分的聚类熵，可以为类别数目的发现提供依据。聚类熵[6]是对类别划分质量进行评价的重要指标，其计算公式如下：

$$\mathrm{En}(k) = \left(\sum_{j=1}^{k} \sum_{i=1}^{n_j} \mathrm{Rel}_d\left(p_i^{(j)}, p_0^{(j)}\right) \right) + \sum_{j=1}^{k} \mathrm{Rel}_d\left(p_0^{(j)}, c_0\right) \quad (1)$$

其中公式右边的第 1 项代表类内熵值，第 2 项代表类间熵值。c_0 表示所有样本的中心，$p_i^{(j)}$ 表示第 j 类的第 i 个样本，$p_0^{(j)}$ 表示第 j 类样本的中心，k 值为聚类的个数，n_j 为第 j 类中的样本个数。样本集合的中心为所有样本向量的均值，因此也可以采用文档向量的形式来表示。取 En 最小时的划分作为层次聚类的最终结果。

本文将相关度 $\mathrm{Rel}_c(C_i,C_j)$ 的大小作为 AHC 算法中类别合并的依据，而不是采用传统上单纯基于词频的相似度作为类别合并指标。下面介绍采用相关度的原因及其具体计算方法。

传统上一般将文档表示为单纯基于词频的向量，并利用向量间的夹角余弦计算文档相似度。即对于任何两个文档向量$\boldsymbol{A}$，$\boldsymbol{B}$，其余弦相似度[7]计算公式如下：

$$\mathrm{sim}(\boldsymbol{A},\boldsymbol{B}) = \frac{\sum_{j=1}^{n} w_{A_j} w_{B_j}}{\sqrt{\sum_{j=1}^{n} w_{A_j}^2} \sqrt{\sum_{j=1}^{n} w_{B_j}^2}} \quad (2)$$

其中 n 为向量 $\boldsymbol{A}$，$\boldsymbol{B}$ 的维数(二者维数相同)，w_{A_j}，w_{B_j} 分别表示向量 $\boldsymbol{A}$，$\boldsymbol{B}$ 在第 j 维上的权值。

由于文档聚类是一种面向自然语言的应用，因此在进行类别划分时往往涉及到一些语义问题，这也是人工标注同类文档的一个重要依据。如果同类文档之间的公共词较多，通过单纯基于词频的相似度计算方法可以达到聚类的目的。但如果同类文档之间的公共词较少，则通过单纯基于词频的相似度计算方法难以将同类文档聚合在一起。通过对同类文档进行内容词分析，发现不同文档的内容词之间往往具有较强的相关性。所以可以计算类别(文档)的相关度，并将类别(文档)相关度大小作为类别合并的依据。本文采用知网中的词概念相关计算模型来计算文档相关度，并利用文档相关度计算类别相关度。知网作为一个知识系统，它所着力要反映的是概念的共性和个性。知网中每个词一般对应多个义项，义项往往由多个义原组成。义原是最基本的、不易于再分割的意义的最小单位。义原使知网中的某些词与其他词产生相关关系，例如知网中和“医生”相关的词有“疗效”、“患者”等。

显然，计算文档之间的相关度离不开词一级的支持。因此本文利用知网作为知识源，对文档中比较重要的特征词进行了必要的扩充，形成文档相关词向量。而两篇文档之间的相关度则定义为传统单纯基于词频的相似度与基于相关词向量的相似度的加权和。具体计算公式为

$$\mathrm{Rel}_d(d_1,d_2) = \alpha \mathrm{sim}_1(d_1,d_2) + \beta \mathrm{sim}_2(d_1,d_2) \quad (3)$$

其中 α,β 为权重因子，且 $\alpha+\beta=1$，本文取 $\alpha=0.6$，$\beta=0.4$。$\mathrm{sim}_1(d_1,d_2)$ 表示文档d_1和d_2采用传统单纯基于词频方式计算的相似度，而 $\mathrm{sim}_2(d_1,d_2)$ 则为d_1和d_2在相关词特征空间上的相似度。$\mathrm{sim}_1(d_1,d_2)$ 和 $\mathrm{sim}_2(d_1,d_2)$ 均采用余弦相似度来进行计算，如式(2)所示。但二者的特征空间不同，前者为传统单纯基于词频的特征空间，构造方法可以参考文献。而后者则为扩充后的相关词空间，各维来自于每个文档的相关词向量。算法 1 描述了文档相关词向量的构造方法。在获得每个文档的相关词向量后，相关词空间的构造方法与传统方法相同。

算法 1　文档相关词向量的构造算法

(1) 对输入的文档 d 进行分词，停用词过滤处理，获得内容词矢量 $\boldsymbol{V}$；

(2) 对 $\boldsymbol{V}$ 进行冗余消除处理，获得唯一词集合 S；

(3) 对 $\boldsymbol{V}$ 进行词频统计，并将频率最大的 $\gamma|S|$ 个词放入集合 S'；

(4) For every word w_i in S'；

(5) 到知网中检索 w_i，获得词 w_i 的记录集合 $R(w_i)$, $(|R(w_i)| \ge 0)$；

(6) For every record r_{ij} in $R(w_i)$；

(7) 查找概念相关词集合(定义 rank=0，即最相关词)，构造集合 C_{ij}；

(8) End for;

(9) For j=0 to $|R(w_i)|-1$；

(10) $I_{ij} = C_{ij} \cap S$；

(11) End for;

(12) Find m,且$|I_{im}| = \max(|I_{ij}|)$, $(0 \le m, j \le |R(w_i)|-1)$；

(13) If $|C_{im} \cap S| > N_\theta$，将 C_{im} 中的所有词保存到 S'' 中;

(14) End for;

(15) 对 S'' 进行词频统计，获得文档相关词向量 $\boldsymbol{R}(d)$。

算法 1 中，$|\cdot|$ 表示集合中元素的个数。经过上述算法的处理，可以获取文档的相关词向量。值得指出的是，在构造文档的相关词向量时，只对文中频率较高的代表词进行了概念扩充。因为根据词汇集聚理论[8]，文档可以包含多个不同的词汇链。每个词汇链一般都有一些频率较高的词。由于词汇链上的词大都是相关词，所以不必对所有词进行扩充。γ 表示从文档 d 中取出的代表词的比例，取经验值为 0.05。知网中每个词一般有多个义项，因此算法中进行了消歧处理。另外如果经过概念扩充后的相关词集合与文档正文的特征词的交集较小，如小于 N_θ (本文取 $N_\theta = 3$)，则不对这个词进行扩充。

2.2 关键词抽取

通过上面所述的方法，可以对文档集合有效地进行聚类。同一类的文档之间具有较大的相关度，不同类的文档之间相关度较小。聚类结束后，还需要提供类别的描述，这样可以帮助用户判断各个类别的文档是否与查询相关，进而缩小用户检索的范围。本文采用关键词作为类别的描述，关键词自动抽取的输入为所有同类文档，输出结果为 5 个关键词，用于类别描述。

短语(phrase)比词(word)信息量更加丰富，更能体现原文的主题，且可以保证一定的可读性。科技文献中作者标注的关键词实际上多数为短语形式。本文以1998年1月份的人民日报作为语料，人工标注了7500条关键词短语。并利用粗集理论在数据泛化和知识约简方面的优势，以其作为训练集进行规则挖掘，获得了94条中文关键词短语的一般构成规则。规则的一般形式为

(1) (a0==“∗”)∧(b0==“∗”)∧(b1==“∗”)∧(b2==“∗”)∧(c0==“∗”)=>d=“∗”

(2) (a0==“∗”)∧(b0==“∗”)∧(b1==“∗”) ∧(c0==“∗”) => d=“∗”

规则(1)判断词性分别为b0,b1,b2的3个词构成的短语在左边界词的词性为a0，右边界词的词性为c0的情况下，是否满足关键词短语的构成规则。规则(2)判断词性分别为b0,b1的两个词构成的短语在左边界词的词性为a0，右边界词的词性为c0的情况下，是否满足关键词短语的构成规则。规则中词性的取值范围根据北大词性标注规范[9]而定，规则右边d的取值为1或者0，表示是否满足关键词短语的词性规则。

系统中抽取出的关键词短语必须满足词性构成规则。关键词的抽取还必须考虑到词的重要性评价问题，即构成短语的词对全文内容的概括能力。在进行词的重要性评价时综合考虑了词的出现频率、位置、与线索词的同现信息、词义支持度等因素。经过上述处理后，系统抽取出 15 个候选关键词短语。然后以知网作为知识源，将某些相似度较大的词，以及那些存在字符串包含关系的词进行冗余消除处理，最后输出 5 个关键词短语作为聚类描述。由于篇幅所限，这种关键词短语抽取技术的细节将另文描述。

3 实验结果和分析

通过实验考察了本文提出的聚类方法及其对信息获取效率的影响。具体做法是分别选取 5 个查询词发出检索请求，然后从返回文档中选取一部分进行聚类处理并提取出不同类别的关键词作为类别描述。测试所用的文档数据来自 Google 的检索结果，聚类结果如表 1 所示。

表 1 聚类结果

Tab.1 Clustering results

序号	查询词	文档总数	聚类后的类别数
1	旅游	80	4
2	搜索引擎	80	6
3	海啸	80	4
4	病毒	80	4
5	生物技术	80	5

通过对同一查询返回的文档进行聚类处理，可以将其自动分为若干个不同的类别。实验中通过对类别内部相关度、类别之间相关度的数值进行对比分析，考察了聚类结果是否较好地遵循了聚类假设。聚类假设由 Jardine 和 Van Rijsbergen[10]首次提出，其基本思想是：文档之间的关联程度表达了文档对请求的相关性，紧密联系的文档属于同一类别，并且都和同一请求相关。因此一个理想的聚类结果应该保证同类文档之间的相关度较强，不同类文档之间的相关度较弱。

表 2 给出了同一类别内部和不同类别之间的相关度以及相似度的对比情况。其中不同类别之间的相关度、相似度的计算方法如定义 3 和定义 5 所示，而同一类别内部的相似度为类别内部任何两个不同文档之间的相似度的均值，同一类别内部的相关度为类别内部任何两个不同文档之间的相关度的均值。可以看出通过采用类别相关度作为聚类合并的指标，使得同类文档之间的距离更接近了，同时也拉大了不同类文档之间的距离。这种方法对于搜索引擎返回结果的后处理是必要的。由于搜索引擎返回的结果来自同一检索词，因此相对于其他类型的聚类数据(如文本分类数据，不同新闻组的数据等)其类别可分性不是特别明显。而基于相关度的方法通过深入到语义一级，可以挖掘出隐藏的类别差异，从而帮助用户迅速找到所需的信息。

为了考察聚类方法对信息获取效率的改善能力，请了 5 个志愿者进行对比测试。具体做法为：对于每次查询返回的

表 2 类别内部与不同类之间的相关度以及相似度的取值对比

Tab.2 The relevance, similarity of intra-cluster and inter-cluster

查询序号	距离指标	同一类别内部	不同类别之间
1	基于概念扩充的相关度	0.4516	0.3726
1	单纯基于词频的相似度	0.3973	0.3502
2	基于概念扩充的相关度	0.5131	0.3481
2	单纯基于词频的相似度	0.3724	0.3166
3	基于概念扩充的相关度	0.645	0.5029
3	单纯基于词频的相似度	0.5063	0.4582
4	基于概念扩充的相关度	0.3544	0.2161
4	单纯基于词频的相似度	0.2495	0.1802
5	基于概念扩充的相关度	0.3872	0.2939
5	单纯基于词频的相似度	0.3176	0.2851

80 个文档，选取其中 2 个文档作为查询对象，事先让用户阅读这 2 个文档的正文，从而了解其主要内容。然后在两种模式下(Google 采用的线性排列模式和聚类后的多类划分模式)，让用户从这 80 篇文档中查找他们曾经阅读过的文档。结果表明通过对检索结果进行聚类处理和关键词自动生成，使用户找到文档的时间大大缩短，如表 3 所示。表 3 中的数据表示对于 10 个文档(5 次查询，每次选取返回结果中的 2 个文档)，5 个志愿者(User 1,User 2,User 3,User 4,User 5)在两种不同模式(A 表示 Google 采用的线性排列模式；B 表示聚类后的多类划分模式)下找到曾经阅读过的文档所用的时间。

实际应用中从搜索引擎返回的结果往往是海量的信息，所以对全部文档进行聚类处理是不现实的。因此可参照文献的思想，先对一部分文档(例如排在前面的文档)进行聚类和关键词生成，然后利用各个类别的关键词作为类别特征，对其余的文档进行分类处理。

表 3　聚类前后的信息获取时间对比(单位：s)

Tab.3 The contrast of information access time before and after clustering

Document ID	User 1		User 2		User 3		User 4		User 5	
	A	B	A	B	A	B	A	B	A	B
1	26	17	35	7	63	23	37	10	23	15
2	14	15	33	21	19	15	15	14	20	24
3	58	23	37	31	40	12	26	25	21	15
4	20	16	26	15	78	15	54	32	23	17
5	35	28	42	21	34	21	32	16	54	32
6	38	12	37	15	36	24	23	7	53	21
7	23	9	18	12	32	14	44	18	32	16
8	18	23	53	16	34	9	36	24	53	25
9	36	21	43	26	26	19	64	31	45	8
10	27	16	54	12	34	20	26	21	64	17

4　结论和展望

实验表明，聚类分析和多文档关键词自动生成技术可以明显提高用户获取有效信息的效率。尽管进行文档聚类、生成关键词等后处理过程也需要时间，但这是由计算机自动完成的，不需要用户的干预。因此采用这种方法减少了用户花费在点击、阅读等人工操作上的时间，减轻了用户的负担。下一步的工作重点是研究如何提高聚类算法的效率。另外一个非常有意义的工作是考察如何利用聚类分析获取的关键词对原始查询进行扩充，从而实现更加精确的查询。

参 考 文 献

[1] Bollacker K D, Lawrence S, Giles C L. Discovering relevant scientific literature on the Web. *IEEE Intelligent Systems*, 2000,15(2): 42−47.

[2] 王爱华，张铭等. PCCS 部分聚类分类：一种快速的 Web 文档聚类方法. 计算机研究与发展, 2001, 38(4): 415−421.

[3] Steinbach M, Karypis G, Kumar V. A comparison of document clustering techniques. Department of Computer Science and Egineering, University of Minnesota, Technical Report #00-034, May, 2000.

[4] 董振东等. 知网.http://www.keenage.com/.

[5] Kaufman L, Rousseeuw P J. Finding Groups in Data: An Introduction to Cluster Analysis. New York: John Wiley & Sons, 1990, chapter 14.

[6] Jung Y. Design and evaluation of clustering criterion for optimal hierarchical agglomerative clustering. [Phd. Thesis], University of Minnesota, September, 2001.

[7] Strehl A, Ghosh J, Mooney R. Impact of similarity measures on web-page clustering. AAAI 2000 Workshop on AI for web Search. Austin, TX, USA, July 2000: 58−64.

[8] Hirst G, Onge D S. Lexical chains as representation of context for the detection and correction of malapropisms. In Christiane Fellbaum, editor, WordNet: An electronic lexical database and some of its applications. Cambridge, MA: The MIT Press, 1997, chapter 13, 305−332.

[9] 俞士汶等. 现代汉语语料库加工规范. http://www. icl.pku. edu. cn/icl_groups/corpus/spec.htm.

[10] Jardine N, van Rijsbergen C J. The use of hierarchic clustering in information retrieval. Information Storage and Retrieval, 1971, 7: 217−240.

刘远超：　男，1971 年生，博士生，讲师，研究方向为自然语言理解、文本挖掘.

王晓龙：　男，1955 年生，博士生导师，教授，研究方向为自然语言处理、人工智能.

刘秉权：　男，1970 年生，博士，副教授，研究方向为自然语言处理、人工智能.

钟彬彬：　女，1982 年生，硕士，研究方向为自然语言处理、人工智能.