基于关键词提取的搜索结果聚类研究[1]

张华平 [1),3)]，秦鹏 [1),2)]，李恒训 [1),2)]，莫倩 [4)]

[1)] (中国科学院计算技术研究所，北京　100190)

[2)]（首都师范大学计算机联合实验室，北京　100037）

[3)] (北京理工大学，北京　100081)

[4)] (北京工商大学，北京　100037)

Email: pipy_zhang@msn.com

摘　要：信息检索的结果往往庞杂，缺乏有效地加工整理，对搜索结果进行聚类是一种普遍的需求，而传统的文本聚类方法不能提供有效的类别标签，且速度较慢，不适用于在线搜索结果的聚类。本文针对性地提出了基于关键词提取的搜索结果聚类算法，基本思想为：结合信息检索的特点，将词频（TF）、词性和互信息等特征进行融合计算，综合实现关键词的提取；最终以筛选出的关键词作为基础特征，实现层次聚类。经实验验证，该方法 P@10 达到 80%，用户满意度达到 85%。实验结果表明，基于关键词提取的搜索结果聚类算法优于目前已知的所有系统。

关键词：关键词提取；搜索结果聚类；信息检索；

Search Result Clustering Based on Keyword Extraction

ZHANG Hua-Ping[2),4)] QIN Peng[1),2)]　LI Heng-Xun[1),2)]　MO Qian[3)]

[1)] (Join Faculty of Computer Scientific Research, Capital Normal University, Beijing 100037)

[2)] (Institution of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

3) (Beijing Technology and Business University, Beijing 100037)

4) (Beijing Institute of Technology, Beijing 100081)

Email: {lqinp, lihx,　zhanghp} @golaxy.cn

Abstract: Web Search results clustering is userd to organize search results which is complicated and poorly organizted , and make it easy for user to browse the results. Web search results is required widly. Traditional clustering techniques are inadequate since they can not generate clusters with highly readable names and they process so slowly that can not meet the requirement. A multi-feature integrated model is developed to evaluate of the keyword , which combines the term frequence, POS, mutual information features together. The improved keyword extraction method takes into account of the feature of search result. According to the experiments, it can be concluded that the method, which P@10 reached 80% and customer satisfaction reached 85%, is better than known system.

Keywords: keyword extraction; search result clustering; information retrieval

1　引言

随着互联网的发展，搜索引擎扮演着越来越重要的角色。目前搜索引擎的使用率为68.0%，在各互联网应用中位列第四。2008年全年搜索引擎用户增长了5100万人，年增长率达到33.6%[1]。然而，当前搜索引擎返回的结果往往庞杂，缺乏有效地加工整理，不能满足用户需求。首先，海量的搜索结果以庞大的线性列表的方式组织，用户只能顺序地浏览并查

[1]本课题得到国家高技术研究发展计划（863 计划） (2007AA01Z438)资助.张华平, 男，1978 年生,博士，副研究员，研究生导师，主要研究方向为中文自然语言处理、信息检索与舆情计算；秦鹏，男，1984 年生，硕士研究生，主要研究方向为中文自然语言处理；李恒训，男，1985 年生，硕士研究生，主要研究方向为中文自然语言处理与信息检索；莫倩，男，1972 年生，博士，副教授，研究生导师，主要研究方向为自然语言处理、信息检索与社会计算。

找所需内容，效率低下；其次，搜索结果歧义混杂，同一搜索词的结果往往多种形式，搜索词本身也存在歧义，例如：“苹果”部分搜索结果和水果相关，另一部分是“苹果”公司的相关动态，搜索引擎将这些内容杂乱得混合到一张列表中返回给搜索该词的用户，用户需要耗费大量的精力才能筛选出想要的信息；更为重要的是，绝大多数普通用户往往很难采用精确的搜索词清晰地表达自己的信息需求，而搜索引擎设计的前提往往假设是用户的查询是精确的，对搜索词的潜在语义扩展（类似于联想功能）是极其必要的。

搜索结果聚类将不同意义的内容以语义聚团的方式组织，解决歧义混杂的问题，方便用户查看信息，带来更好的用户体验；并且搜索结果聚类对查询结果进行全面的分析处理，对同类的结果归类，全面地阐释查询对象，类别标签可以给出查询对象的一组扩展关键词，这些关键词不仅体现了查询对象的特点，而且可以帮助构建新的查询，以找到更全面准确的信息。例如，查询“美国总统大选”，聚类标签是“奥巴马”，“麦凯恩”，用户就可以猜测出和总统大选有关的两个人，获得该事件的一个重要的信息。可见，搜索结果聚类能较好的解决传统搜索引擎呈现搜索结果的诸多不足。

关键词提取，即计算词汇、短语与文章主题内容相关度，按相关度排序返回最能表现文章主题的前 N 个词汇或者短语。本文采用基于关键词提取的方法进行聚类。首先，通过关键词之间的关系可以发现文章之间内在的联系，得到更好的聚类效果；其次，关键词提取过程中使用多特征融合的评价方法，即使文章较短也能提取出高质量的关键词；最后，提取过程中利用互信息计算关键词与搜索词的相关性，并且采用一定的算法防止针对搜索引擎的内容作弊，使得关键词提取方法更适用于检索结果聚类的特征选择。因此，基于关键词提取的检索结果聚类能够得到更好的聚类效果。

2 相关工作

目前已有各种各样的聚类方法，按照聚类方法的特点可以分为：基于文档内容的聚类与基于标签的聚类[2]。

基于文本聚类的方法，主要采用传统的聚类方式，使用搜索引擎返回的标题、摘要替代全文进行聚类进行效率优化。SCATTER／GATHER[3]是第一个 Web 检索结果聚类系统，使用夹角余弦定理计算两个文档的相似性，采用 Fractionation 算法进行聚类。WebCat[4]使用传统的 K-means 算法获得聚类结果。由于标题与摘要信息较少，数据稀疏问题严重，导致聚类结果较差，分析得到的聚类名称不易理解。

目前文本聚类方法已经较为成熟，但是搜索结果聚类不同于普通文档的聚类，不能直接套用传统的聚类方法。首先，普通文档集合中的两篇文档没有必然的联系，而搜索结果集合中的文档主题都是围绕关键词的，文档之间都有一定得关联；其次，由于效率的原因，搜索结果聚类使用的是文章的标题与摘要，包含的信息量少，篇幅短小；最后，部分网站采用搜索引擎优化技术，导致某些垃圾串频率较高。

基于标签的聚类，首先，从文档集中抽取有价值的短语、词组、片段作为标签作为类别名，然后将文档分配给这些类别。SNAKET[5]利用知识库对词语进行评分、选择，选出权重高的词语形成基类，并利用权重较低的词语指导进一步的聚类。词汇相关度聚类利用相关度高的词对指导聚类。SHOC[6]通过后缀词组得到连续的词语作为标签，并利用 SVD 获得层次的类结构。张刚等人[7]采用图的方式，统计文档共现词频，选择词频高的词作为基类，然后通过图的连通关系进行进一步的聚类。

本文属于后一种方法。采用关键词提取的方法进行标签选择，解决传统方法获取类名可读性差的问题。并以关键词作为基础特征进行聚类以及类别合并，获得可读性强的层次化的聚类结果。

3 问题定义

定义相关术语如下：

K：用户输入的搜索关键词；

D={d_i}：D 为搜索结果集合，其中 di 是由搜索结果标题与摘要组成的第 i 个文档。正文的获取速度较慢，不能满足在线聚类的要求，所以本文使用摘要与标题替代全文进行聚类；

C ={DT_1,DT_2,...，DT_i…DTn}：聚类结果集合，其中 DT_i 为第 i 个类别的文档集合；

TD =tag_of(D)：关键词提取的形式化定义，TD 为关键词集合；

C= KBSC (K,D)：KBSC（Keyword Based Search Clustering）聚类算法的形式化定义；

聚类过程：

1)首先，对集合 D 进行关键词提取得到关键词集合 TD，即 TD =tag_of(D)。

2)以关键词作为基础特征进行聚类。即 C= KBSC (K,D)。

4 关键词提取

4.1 关键词提取

4.1.1 候选词提取

系统使用 ICTCLAS 分词程序对搜索引擎返回的元信息进行分词[8]。对分词后的结果进行去停用，过滤一些常见的停用词，例如：“是”、“的”。新版的 ICTCLAS 提供专业词典与用户词典的功能，用户可以添加自己的词典。本文在分词的过程中加入了关键词词典。该词典中收录了大量人工标引的关键词。一些期刊、杂志网站收录了许多人工标注的关键词的文章，并且这些文章都有具体的分类，从学术文章，到科普生活类的文章，包含的词语非常的丰富。本文使用的关键词词典就是这些文章中的关键词。首先，使用采集程序抓取包含这些文章的网页，并抽取其中的关键词、类别等信息，获得 200 万的词典。该词典中存在部分垃圾词汇，主要是抽取过程中带来的噪音，以及部分期刊的关键词是机器自动生成的垃圾词汇。然后，对词典进行人工清洗，从而获得最终 170 万的关键词典。

4.1.2 新词发现

基于词典的关键词最大的不足之处就是不能发现新词，而新词往往包含重要的信息，很可能是当前的热点。例如，搜索“流感”，当前的热点是“猪流感”，词典中未收录该词，则这样重要的信息就会丢失。文献[9]提出的基于局部性原理的有意义串提取方法对新词的识别有较好的效果，能够识别出“猪流感”这样的新词，是解决新词发现的的有效方法。但是，该方法应用在检索结果聚类中会带来新的问题。例如，对于给定的语料，使用该方法可能同时识别出“猪流感”、“猪流感疫情”、“猪流感病毒”等词，这些词往往词义相近或者相同，如果都作为类别的名称，会给类名选择的过程带来新的噪音，使聚类效果下降。各类间不仅意义相近，而且类别中的文档重复率高。本文针对上述问题，对该算法进行改进，加入词语相似性的计算，合并意义相似的词语，使有意串的方法更适用于检索结果聚类。改进后的算法如下：

算法 1：基于局部性原理的有意义串生成算法改进

输入：语料 C，词典 D，重复串频次阈值θ_1，邻接类别 AV 阈值θ_2，参数 l，参数 λ，密

度阈值θ_3，打分阈值θ_4，W(s)表示 s 的字符集合,相似度阈值θ_5；

输出：有意义串集合；

1. 对 C 用词典 D 进行分词
2. 对分词后语料 C 发现特征频次大于 θ 1 的频繁模式，形成集合 FP(C)
3. 对 FP(C) 中每个字符串 S 执行步骤 4 到步骤 9
4. 计算 AV(S)，如果 AV(S)< θ_2，删除 S
5. 将 C 划分称大小为 l 的区域，计算每个区域的密度值
6. 查找密度大于θ_3的连续区域 domain1, domain2 … domainN, 计算各个区域对应的位置方差 $D_1,D_2\ldots D_N$，由 $D_1,D_2\ldots D_N$ 计算局部性度量值 LE(S)
7. 计算 f(S)，如果 f(S)< θ_4，删除 S
8. $\forall S_i \in FP(C)$，如果$W(s)\cap W(s_i)>\theta_5$,删除 S
9. 仍在 FP(C)中的字符串为有意义串，输出有意义串

其中 f（S）表示 S 成为有意义串的分数值。f（S）越大，S 越有可能是有意义串。

4.2 多特征融合的关键词评价方法

4.2.1 TF 特征

TF 表示词语在文档中出现的频率，即词频。一般认为一个词的 TF 越高，则该词在这篇文章中越重要。但是搜索引擎返回的元信息不同于一般性的文章。通常的搜索引擎使用采集器爬取互联网上的信息，不仅收录了有用的信息，也抓取了大量垃圾信息。例如：一些网站为了进行搜索引擎优化，在一个页面中包含大量的重复的词。当用户搜索该词时就会返回该页的信息，而这样的信息对用户是没有用的。令文章的词总数为 C_{total}，P_i=TF/C_{total},当 P_i >P 时（其中 P 为阈值，一般取值 0.75），则认为该文章包含垃圾信息过多，文章重要度低。

$$TF_{new} = \begin{cases} \text{TF} & \text{TF}/C_{total} > \text{P} \\ 0 & \text{TF}/C_{total} \le \text{P} \end{cases}$$

4.2.2 词性特征

根据汉语的特点，发现关键词一般都是名词(n)与动词(v)，还包含少量的形容词与副词。介词，助词等一般不能表述具体的意义。而名词中的人名(nr)、地名(nt)，机构名(ns)等更有可能成为关键词。所以本文使用词性进行关键词权重的调整。令集合 A={nr,nt,ns,v}，t 为关

键词，P_{weight} 为词的词性权重，p 为候选关键词词性，T 关键词集合，P 为关键词对应词性权重集合，a、b、c 为可调节变量，一般取值分别为 3、2 和 1。如果 $\forall$ p∈ A，当 p=nr 时 P_{weight} = TF_{new}*a；p=（ns|nt）时，P_{weight} = TF_{new}*b； 当 p∈ $n \cap$ p∉ A 或者 p=v 时，P_{weight} = TF_{new}*c，则将 t 加入集合 T 中，将 P_{weight} 加入集合 P 中。最终集合 T 为过滤后的关键词集合，P 为对应词的词性权重。

4.2.3 互信息特征

检索词是检索结果的核心主题，在检索词周围出现的词更有可能成为关键词，例如，搜索“流感”，在检索结果中“流感”周围出现的词频较高的为“甲型 H1N1”，“疑似病例”等。这些词都能较好的体现搜索结果的主题，更适合做关键词，所以关键词的权重大小应该体现检索词与关键词之间的关系。本文使用互信息表示两个词的依赖关系。据文献[10]，互信息定义如下：

$$MI = Log\frac{p(query,term)}{p(query)*p(term)}$$

其中 p(query,term)表示，关键词与查询词在一句话中的共现频率。p(query)，p(term)表示查询词、关键词分别出现的频率。

4.2.4 多特征融合

TF，词性，互信息是评价关键词重要程度的三种重要特征。根据三种特征在关键词中评价中重要程度的不同，本文给出下列公式：

$$w_i = k_1 TF_{new} + k_2 P_{weight} + k_3 MI$$

其中 k_i 为可调节参数，一般取值为 0、1 和 3。在关键词抽取过程中，通过将 TF 特征，词性特征及互信息特征融合到统一模型的多特征融合模型进行评价。

5 搜索结果在线聚类算法

搜索引擎为了用户能够得到更好的用户体验，检索结果一般按照相关度进行排序，越相关的文档越排名靠前。所以，TopN 的文档能够代表用户检索的意图，包含大部分用户希望检索文档的类型。考虑到检索结果聚类的实时性要求，本文使用 TopN 个文档进行聚类，分析获取类名。然后，使用分类的方法，将文档划分到已有的类别中。基于上述思想，本文提出基于关键词的检索结果聚类方法 KBSC，具体算法如算法 2 所示：

算法 2：检索结果聚类 KBSC

输入： 检索词 K，文档集合 D

输出： 类集合 DT

1. 从集合 D 中选取 TopN 文档形成文档集合 DD

2. 对 $\forall\ d_i \in$ D 执行 tag_of(d_i)得到集合 T{TD_1,TD_2,…,TD_i , TD_n},其中 TD_i ={t_1,t_2,…,t_j ,t_n}表示文档 d_i 的关键词集合,t_i 为第 i 个关键词

3. 对于 $t_j \in TD_i$ (0<i<|T|, 0<j<|TD_i |)，如果 $\exists t_k \in TDD$ ， t_k 的值与 t_j 相等，则 t_k .weight += t_j .weight，否则将 t_j 加入集合 TDD 中

4. 对于每一个 $t_i \in$ TDD（0<i<|TDD|）,如果 $t_i \in$ di，则将文档 d_i 加入集合 DT_i 中，形成集合 DT= {DT_1,DT_2,..., DT_i …DT_n}

5. 对于所有 0<i<| DT_i |， 0<j< DT_j ，如果 $DT_i \cap DT_j / \min(|DT_i|,|DT_j|) > p_{dt}$ （ p_{dt} =0.75），执行 6，如果 $DT_i \cap DT_j / \max(|DT_i|,|DT_j|) > p_{mdt}$ （ p_{mdt} =0.9），执行 7，否则执行 8

6. 将 t_i 与 t_j 合并，形成新的 t_i ,删除集合 DT_j

7. $DT_{\min}$ = DT_i ， $DT_{\max} = DT_j$ ； 如果|DT_i |<|DT_j |，则 $DT_{\min}$ = DT_j ， $DT_{\max} = DT_i$ ； $DT_{\min}$ 为 $DT_{\max}$ 的子类别

8. 从集合 DT 中选取 TopN 形成集合 C ={DT_1,DT_2,..., DT_i …DT_n}为最终类别

6 实验结果及分析

6.1 实验设计

本文采用用户评价的方法，对类名进行评分。使用 30 个不同类型的查询，如表 1 所示，查询的类型包括：歧义查询词、热点词和一般关键词。热点词使用搜索引擎中热门搜索词中的前 10 个词，热门搜索词是一段时间内用户检索词数次数较多的词，能够较好的反应用户的意图，并且这些词的类型丰富，包含新闻、娱乐、军事、体育等等。

表 1 不同类型关键词列表

Tab.1 List of queries with different types

类型	举例
歧义查询词	苹果，sun，甲壳虫，…
热点词	开心网，我的兄弟叫顺溜，nba,…

一般关键词　　总统大选，世界杯，信息检索，…

本文使用“百度”搜索引擎作为数据源，每个查询词采集 200 条搜索结果进行聚类。

6.2　多特征融合方法评测

采用表 1 构造的查询词对多特征融合方法进行测试，测试词性 POS 与互信息 MI 对多特征融合公式的影响，希望通过实验验证多特征融合公式的有效性，

表 2 不同特征排序算法评价

Tab.2 Evaluation of feature combination

	P@3	P@5	P@7	P@10
TF+MI	77%	80%	73%	68%
TF+POS	83%	84%	82%	79%
TF+POS+MI	86%	86%	84%	80%

从表 2 可以看出，词性对准确率的影响较大，因为词性能够过滤介词，助词等一般不能表述具体意义的高频词汇，调整名词、动词的权重；互信息对准确率的影响较小。但是，互信息能够发现重要的低频类名，例如：搜索“总统大选”，搜索结果大多数都是与“美国总统大选”相关内容，“伊朗总统大选”、“俄罗斯总统大选”等信息较少、排名靠后，但是这些词语与关键词之间的互信息值较大，通过互信息特征能够将这些类别的排名提前，提高用户满意度。

因此，多特征融合的方法能够更好的描述关键词，获得更高准确率。

6.3　与已有系统对比实验

使用表 1 构造的查询词，将本文提出的聚类方法产生的聚类结果与 vivisimo 聚类结果进行比较。Vivisimo 是一个元搜索引擎，以层次化聚类的形式展现搜索结果。在 2001 至 2003 年被搜索引擎观察的专家评为“最好的元搜索引擎”，是国际上公认的搜索结果聚类软件。

测试对象是中科院 40 位不同专业的学生、老师。如图 2 所示，其中有 65%的人认为 KBSC 的聚类结果优于 vivisimo，35%人认为 vivisimo 聚类效果更佳。图 1 是用关键词“总统大选”进行的对比，左边为 KBSC 聚类结果，右边是 vivisimo 聚类结果。

图 1 关键词“总统大选”，KBSC(左)与 vivisimo(右)聚类结果

Fig.1 Clustering of Keyword” 总统大选”, KBSC(Left) and vivisimo(Right)

同时，对 KBSC 产生的聚类结果进行了用户满意度调查，如图 3 所示，其中 55%的用户对聚类结果较为满意，30%的用户认为聚类效果一般，15%的认为聚类效果较差。据调查，聚类结果存在的主要问题是，类别的重复。例如，图 1 中，类别“美国总统大选”子类别中又包含了“选举”。从图 1 可以看出，Vivisimo 也存在同样的问题。

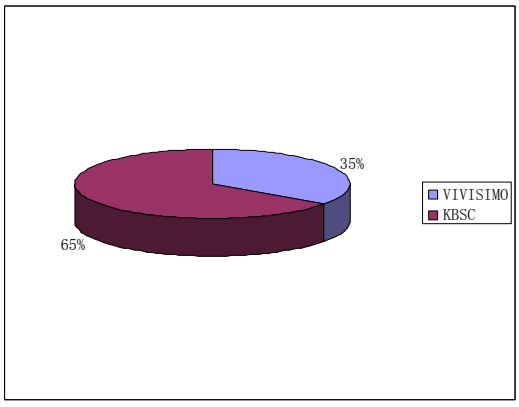

图 2 KBSC 与 VIVISIMO 对比

Fig.2 the comparison of KBSC and vivisimo

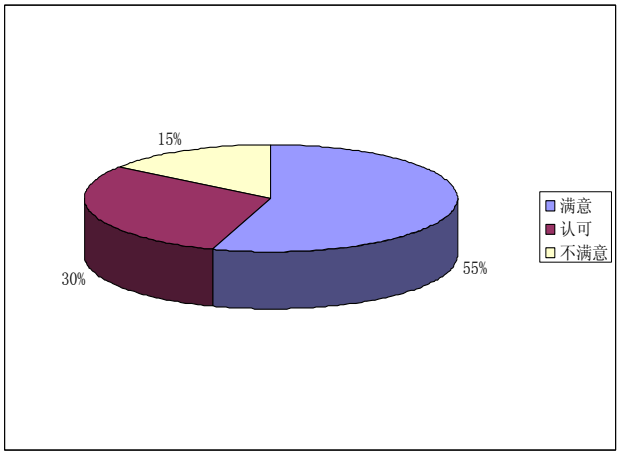

图 3　KBSC 用户满意度调查

Fig.3 The customer satisfaction survery of KBSC

7　总结及下一步工作

本文将关键词提取技术与检索结果的特点想结合，提出将词频，词性和互信息等特征相结合的多特征融合评价方法抽取关键词。并以关键词作为基础特征进行层次化结果聚类。实验表明，本文的的方法能够提取出可读性强、有意义的聚类标签，并且 P @10 达到 80%，用户满意度达到 85%，聚类效果优于商业系统 vivisimo。并且 KBSC 应用于中科天玑舆情监测系统取得了较好的效果，具体信息可从中科天玑网站获取(http://www.golaxy.cn)。KBSC 聚类结果存在一定的重复，降低用户满意度，在接下来的工作中，主要针对类别重复的判定与合并进行优化。

参　考　文　献

[1] CNNIC. 《第 23 次中国互联网络发展状况统计报告》. 北京，2009 年 1 月

[2] Hiroyuki Toda, Ryoji Kataoka. A search result clustering method using informatively named entities. Proceedings of the 7th annual ACM international workshop on Web Information and Data Management. 2005, 81-86

[3] Marti A. Hearst, Jan O. Pedersen. Reexamining the cluster hypothesis: Scatter/gather on retrieval-results.Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval. 1996, 76-84

[4] F. Giannoni, M. Nanni, D. Peclreschi. Webcat: Automatic categorization of Web search results. SEBD03, Cetraro, Italy, 2003

[5] P.Ferragina, A. Gulli, A personalized search engine based on Web-snippet hierarchical clustering, Proceedings of Special interest tracks and posters of the 14th international conference on World Wide Web International World Wide Web. 2005, 801-810

[6] D. Zhang, Y. Dong. Semantic, hierarchical, online clustering of Web search results. The 6th Asia Pacific Web Conference (APWEB).2004, 19-78

[7] 张刚,刘悦,郭嘉峰,程学旗.一种层次化的检索结果聚类方法. 计算机研究与发展，2008,45(3):542-547

[8] 刘群，张华平，俞鸿魁等. 基于层次隐马模型的汉语语法分析. 计算机研究与发展，2004.8

[9] 黄玉兰，龚才春，许洪波，程学旗. 基于局部性原理的有意义串提取方法. 第四届全国信息检索与内容安全学术会议论文集，2008.11

[10] Jing Zhao, Jing He. Learning toGenerate Labels for Organizing Search Results from a DomainSpecified

Corpus. Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence.IEEE Computer Society, 2006,390-396