基于主题短语的搜索引擎结果聚类①

索红光 孙珊珊 王玉伟 梁玉环

(中国石油大学(华东) 计算机与通信工程学院 山东 东营 257061)

摘 要: 为了解决搜索引擎检索结果中的主题混杂现象，帮助用户快速准确地定位到有价值的信息，提出基于主题短语的搜索引擎结果聚类方法。首先从检索结果中提取查询词并与相邻词语组成主题短语，建立包含高频独立词语及主题短语的混合向量空间模型，同时引入同义词词林对特征项进行语义扩充，最后采用改进的 k-means 聚类算法对搜索结果进行聚类，并为各个类别提取类别标签。实验结果表明，该算法能有效提高聚类结果的准确率。

关键词: 搜索引擎；聚类；主题短语；同义词扩展；k-means 算法

Subject Phrase-Based Clustering Algorithm for Search Engine Results

SUO Hong-Guang, SUN Shan-Shan, WANG Yu-Wei, LIANG Yu-Huan

(School of Computer and Communication Engineering, China University of Petroleum, Dongying 257061, China)

Abstract: To solve the problem of mixed subjects returned by search engine results, a new subject phrases clustering algorithm is presented to help locate the valuable results that the users really need. The algorithm firstly extractes some subject phrases from the search results. Then, the vector space model is built. Finally, the results are clustered by the improved k-means algorithm. The algorithm was tested and validated by the experiments.

Keywords: search engine; cluster; subject phrase; synonyms extension; k-means

1 引言

现有搜索引擎主要以查询词为基本索引单位，当某些缺乏经验的用户进行查询时，检索到的结果往往比查询本身的语义更加广泛，比如查询 java，在百度返回的前 20 个结果中有 java 培训、java 技术、java 软件或游戏等多方面的内容。这些结果虽然符合目前搜索引擎基于关键词词形匹配的搜索要求，但表达的主题却完全不同，所以在搜索过程中不得不花费大量时间对候选结果进行筛选。因此，如何从语义层次上合理的组织结果集，使同一主题的结果自动归为一类成为十分关键的问题。解决这一问题的有效方法是对搜索引擎的检索结果进行再处理即聚类。聚类是一种根据对象之间的相似性关系对各个对象进行自动归类的方法。

近年来已经有人提出了几种搜索结果的聚类方法，比如文献[1]是采用后缀树聚类(suffix tree clustering,STC)算法，把文档看作有顺序的词串而不仅仅是一个单词的集合,并通过文档共现的词串来作为文档相似度测量的基础。如两个文档共有至少一个词串,则将它们合并为一个基类。后缀树算法只是简单将共享相同短语的文档归为一类,而没有考虑到自然语言中存在的同义词和多义词现象,从而使得聚类不够准确和完整。文献[2]采用关联规则，通过寻找频繁项集发现各网页间的关系，算法最大的问题在于需要下载整个网页进行分析,因而耗时较长,效率不高。而著名的基于概念的 Lingo[3]聚类算法，首先从返回结果中确定概念文档，从频繁出现的相联系的特征词形成一定的概念文档作为类标记。其中要对特征词-文档矩

① 收稿时间:2009-06-25

阵进行奇异值分解，这是一个耗时的过程。

基于查询词匹配的搜索结果中包含大量查询词，这些词往往语义广泛对网页主题没有很好的区分作用，而与查询词相关联的短语语义更加具体，能够更加准确的表达网页的主题。基于上述问题，本文对检索结果进行再处理，将检索结果中出现的查询词与前后相邻词语分别组成主题短语，并建立包含高频独立词语及主题短语的混合向量空间模型，同时为了提高词频统计的准确性，引入同义词词林对特征项进行扩充。考虑到搜索引擎对检索速度的要求，本文利用改进的 k-means 算法对查询结果聚类，并为各个类别设置类别标签，用户可以根据标签的提示选择相应类别以缩小查询范围，从而快速找到需要的信息。这种方式简便灵活，且易于实现。

2　主题短语的抽取及向量空间表示

针对特定查询词获取返回的搜索引擎查询结果列表，由于结果集数量太大且排在后面的网页一般质量较低，因此只选取一定数目排名靠前的网页进行再处理。考虑到获取网页正文需要一定时间，影响检索速度，本文选取检索结果列表中的网页片段作为研究对象，包括网页的标题，摘要，URL。去除重复、无用链接后，采用正则表达式或 HTMLParser 提取网页片段。

将结果的标题、文档片断和 url 按照标点符号分成一个个单独的字符串或分句，然后利用中科院的分词系统对每个分句进行分词，并去除停用词。对各分句中出现的查询词分别与其前后相邻的一个词语组成主题短语，由于主题短语是在各分句内部组合，所以保证了短语语义的相对完整性。采用 tf-idf 权重计算法计算各词语及主题短语的权重 w_{ij}：

$$w_{ij} = tf_{ij} * \log(N / df_i) \tag{1}$$

其中 tf_{ij} 是词或短语在文档 d_i 中出现的次数，df_i 是词或短语在多少个文档中出现，N 代表总的文档个数。

在搜索引擎返回的结果中会包含一些同义词或近义词，如笔记本和电脑，在传统基于词形的词频统计方法中这些词的词频将各自单独统计，因此在词频统计过程中引入同义词词林，通过同义词词林，一些同义或近义的词将被归为一类，提高了词频统计的准确性。另外认为出现在标题中的词语或主题短语相对更为重要，因此对其进行加权，加权系数为 3，摘要中出现的特征加权系数为 2，url 中为 1，因此权重公式变为：

$$w_{ij} = \lambda * tf_{ij} * \log(N / df_i) \tag{2}$$

其中 λ 表示加权系数。

根据统计的权重，建立各个网页的混合特征向量，该特征向量中不仅包括传统方法中的高频词还加入了本文提出的主题短语，混合特征表示方法能够更加准确的提取网页的主题语义。$d_i = ((t_{i1}, w_{i1}), (t_{i2}, w_{i2})..., (t_{in}, w_{in}))$，其中 d_i 表示第 i 个网页，tf_{ij} 表示该网页中的第 j 个词语或主题短语，w_{ij} 表示的权重。

3　聚类过程

3.1 文档相似度计算

文本聚类的相似度计算本文采用夹角余弦法，夹角越小，相似度越大。定义两篇文本 d_i, d_j 的相似度如下式：

$$Sim(d_i, d_j) = Cos(d_i, d_j) = \frac{d_i \cdot d_j}{\| d_i \| \| d_j \|} \tag{3}$$

为了减小不同长度的文本对于计算文本相似度的影响，对每个文本向量进行归一化，即

$$d_i = d_i / \| d_i \| = \frac{(w_{i1}, w_{i2}, ... w_{in})}{\sqrt{w_{i1}^2 + w_{i2}^2 + ... + w_{in}^2}} \tag{4}$$

其中 n 为特征向量的维数。由公式(4)可得 $\| d_i \| = 1$，其余弦相似度即为两文本向量的点积

$$Sim(d_i, d_j) = d_i \cdot d_j \tag{5}$$

3.2 初始类生成

搜索引擎作为网络检索工具，对检索速度具有很高要求，它们希望能够在最短的时间内返回给用户最满意的结果。k-means 聚类算法实现简单，收敛速度快，其时间复杂度与对象数目成线性关系，对处理大数据集具有良好的可伸缩性和高效性，在实际中得到广泛应用。基于此本文采用 k-means 算法对检索结果进行聚类。

k-means 方法是基于划分的聚类方法，基本思想是：事先给定聚类数目 k，随机选择 k 个文本作为初始类中心，然后计算各个类中心与每个文本的相似度，将文本赋予最相似的类，然后重新计算类中心。不断迭代，直到目标函数收敛。

算法中聚类个数 k 及初始划分需要人为设定，所以对聚类效果影响较大。本文采用文献[4]中的算法生成初始类，通过实验发现在前 20 项返回结果中基本已经包括了查询涉及的主要主题，因此取前 20 项结果文档进行初始类的生成。其基本思想是：初始选定的 k 个文档之间的相似度应尽量的小，即文档间的距离应尽量远。在文档集中任意选取一个文档作为第一个类，计算其余文档与该文档的相似度，选取相似度最小的文档作为第二个类，计算其余文档与前两个文档的相似度，取与这两个文档相似度之和最小的文档作为第三个类，依次类推直到得到的相似度之和的最小值超过一定阈值。

3.3 采用改进的 k-means 算法对搜索结果聚类

由于 k-means 算法容易陷入局部最值，本文对 k-means 算法进行二次分区调整[5]，调整完毕后，重新进行 k-means 启发式的搜索过程。调整过程的基本思路为：当 k-means 算法陷入局部极值时，将聚类分区中的文本向量与 k-means 算法生成的其它中心比较，如果将该点从当前类移动至其它类时，会使目标函数的改变满足一定条件，则移动该向量。对于划分后的类集合 C，目标函数为 k 个类的类内相似度之和

$$Q(C)=\sum_{j=1}^{k}q(C_j)=\sum_{j=1}^{k}\sum_{d\in C_j}d^T Z(C_j) \qquad (6)$$

其中，j=1,2,...,k 表示聚类产生的 k 个类，$q(C_j)$ 表示第 j 个类的类内相似度，$Z(C_j)$ 表示第 j 个类的类中心，将一个向量 y 从其所属类 C_i 移至类 C_j 时目标函数的变化记为：

$$\Delta Q=Q(C^{(t+1)})-Q(C^{(t)}) \qquad (7)$$

k-means 算法改进后的详细步骤如下：

Step1：利用传统 k-means 算法得到类的集合。

Step2：由公式(6)(7)计算文本向量 $d\in D$ 从类 C_i (d 所在的类)移动至类 C_j 时的 ΔQ_j $(1\le j\le k, j\ne i)$，若 $(\max\Delta Q_j>\varepsilon'(\varepsilon'>0))$，则将相应文本移动到当前类 C_j 中，记录当前 C_i、C_j 为发生改变的类，否则不作改变。重复 Step2，直到遍历完所有文本，其中每个文本只遍历一次。

Step3：由 Step2 得到新的类分区集合记为 $C^{(g+1)}$，更新经过移动修改的簇的中心向量 $Z_j^{(g+1)}$，$1\le j\le k$，其中 g 为算法调整的次数。

Step4：如果所有 $\Delta Q<\varepsilon'$ $(\varepsilon'>0$，为判断终止的阈值)，终止算法，输出聚类最终类的集合 C；否则 $g=g+1$，转到 Step1。

通过上述聚类过程，某查询下搜索引擎返回的结果被自动归为 k 类，每个类别内的文档具有相对较高的相似度。

4 类别标签的提取

为了更加准确的概括类内主题，方便用户的查询，对产生的聚类提取类别标签。目前常见的搜索引擎聚类技术[3]都是采用传统的提取词语或整个句子做类别标签。词语简短但往往包含多种语义对类别内容的提示效果差，句子过长对类别的描述性差。因此本文采用混合方法，提取类内高频主题短语及类中心的高频词作为类别标签。其中的主题短语意义完整消除了单纯使用词语时的歧义，高频词由于来自聚类产生的类中心，对类别有较准确的概括。通过类别标签的提示，用户可以快速进入需要的类别，有效提高了用户查询速度及搜索引擎友好性。

5 实验及结果评价

采用 baidu 作为源搜索引擎，利用 10 个不同的查询进行实验，从每个查询的返回结果中选取前 100 项作为实验数据，采用中科院 ICTCLAS 中文分词系统对查询结果进行分词处理，然后提取特征项建立向量空间模型，利用改进的 k-means 算法对其聚类，最后分别比较该算法产生的类别与人工分类得到的类别，并采用平均查准率 $\overline{P}$ 作为评价指标，$\overline{P}=\sum_{j=1}^{k}P_jP_j$，其中 $j=1,2,...,k$ 是对某查询结果聚类产生的 k 个类别，P_j 表示第 j 类的查准率，即聚类类别中与人工分类一致的文本数占人工分类数的比例。选取五对实验结果如表 1 所示：

表 1 聚类结果

查询词	文档总数	类别个数	平均查准率(%)
java	100	5	0.721
熊猫	100	4	0.696
毕业生电影	100	4	0.690
苹果	100	5	0.713
病毒	100	4	0.707

通过表 1 可以看出，5 次查询请求下的平均查准率都在 70%左右，说明聚类得到的结果中有 70%的文本

可以得到准确归类，通过该聚类算法可以较准确的引导用户对查询结果的检索和过滤，基本达到预期目标。

此外，实验中对每个查询下具体的实验结果做了分析，表 2 给出了对“毕业生电影”查询结果聚类产生的类别标签、各类别中的文档数，并通过与人工分类的比较，得出各个类别的查准率。

表 2 “毕业生电影”实验结果

类别标签	聚类产生文档数	人工分类文档数	一致文档数	查准率
毕业生电影、演员、影评	34	28	17	0.61
下载、免费、毕业生视频	26	35	22	0.63
毕业生插曲、歌词	29	24	15	0.67
毕业生、院校、就业	11	13	11	0.85

通过对表 2 中实验数据进行分析得出以下结论：由于第四个类别与其他类别的特征差距比较大，很容易就与其他三个类别区分开来,因此它与人工分类的情况非常吻合。而第一类别与第二类别中的很多文档的内容非常相近，因此聚类的结果与人工分类的结果有一定差距。

针对特征项的选取问题，分别对 50 条“熊猫”的检索结果以及 50 条“毕业生电影”的检索结果进行了本文方法与简单分词之间的对比实验，简单分词作为特征项是指只对检索结果进行分词、过滤、去除无用词，不进行相邻词组的组合。实验结果如下表 3 所示：

表 3 特征项提取方式的比较

	基于简单分词的特征项个数	基于主题短语的特征项个数
熊猫	160	145
毕业生电影	175	150

通过表 3 的实验数据可以看出，采用主题短语与高频词语相结合的特征提取方法不但没有增加特征项数目，反而比单纯提取高频词语的特征项数目有所减少，可见采用主题短语与词语相结合的特征表示方法并没有增加计算量。

为了验证主题短语与高频词相结合的特征提取方式对聚类的影响，实验对简单分词和混合特征提取方式下的聚类结果作了比较，平均查准率如图 1 所示：

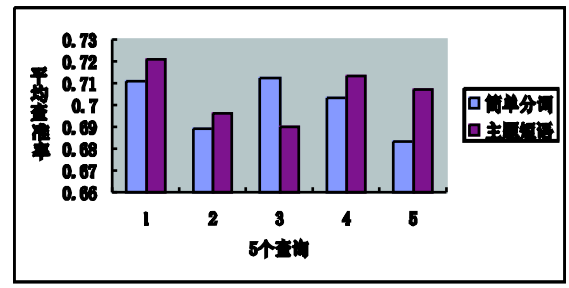

图 1 两种特征抽取方式的聚类结果比较

图 1 中横坐标表示表 1 中列出的五个查询，纵坐标表示平均查准率。从图中可以看出，其中有四个查询在采用混合特征提取方式下的聚类平均查准率要高于简单分词下的平均查准率。说明采用主题短语与高频词相结合的特征提取方式有利于准确挖掘查询结果中的主题语义，提高了搜索引擎返回结果的聚类精度。

6 结语

综上所述，本文分析了当前搜索引擎结果聚类算法的优缺点，对传统基于词语的特征提取方式进行了改进，在提取部分高频词语的基础上，从查询词出现的语义环境入手，提取查询词相邻词语与查询词组成主题短语，并作为特征项的组成。在特征项权重统计中引入同义词词林对特征项进行扩充，避免了同义词和多义词对权重计算准确性的影响。最后利用改进的 k-means 聚类算法对搜索返回的结果进行聚类，提取类别标签。该方法提高了用户查询的准确性，通过实验验证了算法的有效性。

参考文献

1 Zamir O, Etzioni O. Grouper: A dynamic clustering interface to web search results. Proc. of the Eighth International World Wide Conference(WWW8) Toronto，Canada: Elsevier Science. 1999. 1361－1374.

2 宋春芳,石冰.一种基于关联规则的搜索引擎结果聚类算法.山东大学学报, 2006,41(3):68－72.

3 李红梅,丁振国,等.搜索引擎中的聚类浏览技术.中文信息学报, 2008,5(3):56－63.

4 陈晓平,许卓明. WWW 上搜索引擎返回结果的模糊聚类研究[硕士学位论文].南京:河海大学, 2002.

5 索红光,王玉伟.一种用于文本聚类的改进k-means算法.山东大学学报(理学版), 2008,43(1):60－64.