# 基于标签共现的查询扩展研究*

晋松 林鸿飞 苏绥
大连理工大学计算机科学与工程系 大连 116024
E-mail: dutjinsong@gmail.com

**摘 要**：传统的查询扩展方法忽略了查询词与扩展词间的语义关联。随着 Web 2.0 的发展，folksonomy 为网络提供了大量的社会化标注信息。作为 folksonomy 的核心，标签不仅可以高质量描述信息资源的内容和主题，并且标注相同信息资源的标签之间还存在着一定的语义关联。依据标签的这种特性，本文提出了一种基于标签共现的方法来计算标签之间的语义关联程度，通过计算标签与查询词的共现度，可以获得与原始查询语义相关的扩展标签。实验结果表明：与未进行查询扩展时相比，采用本方法进行扩展后，平均准确率提高了 16.7%；这说明本方法选择的扩展标签与原始查询所表征的主题具有较高的语义相关性，folksonomy 的标签资源可以为查询扩展提供丰富准确的扩展词来源。

关键词：folksonomy，查询扩展，标签，共现

# A Folksonomy Tag Co-Occurrence Based Query Expansion Approach for Information Retrieval

Jin Song, Lin Hongfei, Su Sui
Department of Computer Science and Engineering, Dalian University of Technology, Dalian 116024
E-mail: dutjinsong@gmail.com

**Abstract:** Traditional query expansion approaches do not take into account the semantic relationship between the original query terms and expansion terms. Folksonomy is a social service in Web 2.0, which provides a large amount of social annotations. As the core of folksonomy, tags are high quality descriptors of the information contents and topics. Moreover, different tags describing the same information resource are semantically related to some extent. In this paper, we propose a query expansion approach that utilizes the tag co-occurrence information to select the most appropriate expansion terms. Experimental results show that our tag co-occurrence-based query expansion approach consistently improves retrieval performance, when compared with non-expansion method. This indicates that the expansion terms we selected are semantically related to the original query, and tags of folksonomy will be a promising resource for such tasks.

**Keywords:** folksonomy, query expansion, tag, co-occurrence.

## 1 引言

在搜索引擎等实际的信息检索应用中，用户提交的查询请求往往不能准确全面的反应出用户需求，这就会引起信息迷向、信息过载和词不匹配等问题，对检索性能有比较严重的负面影响。如何使查询能够准确的反应用户需求成为信息检索领域中一个重要的研究课题。

查询扩展是解决用户查询模糊性的有效技术手段，它以用户原始的查询为基础，将与原查询相关的词语或概念添加到原查询中，以提供更多有利于判断文档相关性的信息。查询扩展的核心问题是如何选取与原始查询相关度高的扩展词。

当今的 Web2.0 时代，伴随着社会书签类站点的迅速兴起，一种新的组织、管理信息方式也随之诞生，它被称为大众分类法（folksonomy）。用户可将感兴趣的网页收藏到社会书签站点中，

*基金项目：国家自然科学基金资助项目（编号：60373095，60673039）和国家 863 高科技计划资助项目（编号：2006AA01Z151）。

如 del.icio.us[1]，这些网站还允许用户使用自定义标签对文章、图片、视频、声音等资源进行描述，并利用这些用户标签完成信息资源的分类、组织、检索，以便日后查找。

本文通过考察社会书签站点的标签特点，提出了一种基于词共现计算标签相似度的方法。运用该方法从 del.icio.us 网站上提取出语义相关的标签，并利用这些标签进行了查询扩展实验。在 TREC 数据集上的测试结果表明，这种基于标签共现的查询扩展方法可以有效的改善检索性能，其平均准确率提高了 16.7%。实验结果说明该方法可以有效的提取语义相关的标签，这为扩展词的选取提供了一个新的语义来源。

本文的组织结构如下：第二节对查询扩展以及大众分类法（folksonomy）的研究现状进行介绍和分析；第三节对本文提出的计算标签相似度的方法进行详细介绍，阐述语义相关标签的获取方法，并给出了扩展标签的权重计算公式；第四节主要介绍实验目的、所用到的数据集以及结果评测方法；第五节对基于标签共现方法所选取的标签进行分析，并且对查询扩展的实验结果进行分析；最后，第六节总结本文的研究工作并对下一步研究方向进行展望。

## 2 相关领域的研究现状

### 2.1 查询扩展

目前，查询扩展技术可以分为两大类：基于全局语料集分析的方法（简称全局分析方法）和基于局部文档集分析的方法（简称局部分析方法）。

常见的全局分析方法包括 LSI(Latent semantic indexing)、基于词之间相似性词典的方法[1]和 Phrasefinder 方法[2]等。全局分析的优势是可以最大限度地探求词间关系，并在词典建立之后以较高的效率进行查询扩展。但是，当文档集合非常大时，建立全局的词关系词典在时间和空间上往往是不可行的，并且在文档集合改变后的更新代价巨大。因此，近期的查询扩展研究主要集中在基于局部文档集的分析上。

局部分析方法是利用初次检索得到的与原查询最相关的 N 篇文章作为扩展用词的来源。目前，流行的局部分析方法主要是局部反馈(local feedback)[3]，也称为伪相关反馈(pseudo feedback)。Xu 和 Croft 提出来的局部上下文分析方法[4]是利用全局分析的词共同出现频率的思想避免了向原查询加入不相关的词。实验表明，局部上下文分析方法的检索效果明显优于传统的全局分析和局部分析方法。但是，当初次查询后排在前面的文档与原查询相关度不大时，局部分析会把大量无关的词加入查询，从而严重降低查询精度，甚至低于不做扩展优化的情形。

### 2.2 大众分类法和标签

美国信息架构专家 Thomas Vander Wal 于 2004 年 8 月首次提出 folksonomy 这一名词并通过 Gene Smith 的博客传播到全世界[5]。Folksonomy 是由 Folk 和 Taxonomy 两个词合成而来，含义是“由大众的一致意见而产生的基于用户的分类体系”， 中文翻译为“大众分类”、“自由分类”等。Laura Gordon[6]和 Cameron Marlow[7]等以 del.icio.us 等网络书签网站为例，提出标签（tag）是用户用于描述某个信息资源所使用的字、词或短语，它是大众分类法的一个核心。

大众分类法中的标签对资源的主题和内容具有较强的描述性。Xu[8]的研究表明利用标签的这种特性能够有效的提高网络搜索性能。Al-Khalifa[9]等在分析标签含有的语义信息时发现，社会化标签比搜索引擎中的关键字所包含的语义信息还要丰富。Peter Mika[10]和 Wu[11]等的研究发现通过分析标签与资源之间的标注关系可以获得标签的潜在语义信息。以往的研究表明，标签

[1] http://del.icio.us

不仅包含资源的显性知识，同时也蕴含着资源的隐性知识，相同的标签还能够聚合整个信息空间中的所有相似内容，实现资源的共享。

## 3 基于标签共现的查询扩展方法

### 3.1 相似标签的选取方法

大众分类法中，一个资源可以被多个标签进行标注，这些标签之间存在一定的语义关系。通过 del.icio.us 网站的 URL 查询功能，可以得到一个网站使用频度高的一些标签。表 3.1 列出了“google”“ technorati”等六个网站被标注频度最高的 5 个标签。以“google”网站为例，“google”、“searchengine”、“engine”、“web”、“search” 是使用频率较高的几个标签，它们从不同方面对该网站的主题进行描述和概括，并且这些标签之间存在着语义相关性。

根据这点特性，本文提出以下假设：标记同一资源的标签之间存在一定的语义关联。

表3.1 “search” 标签的搜索结果

| 网站 | 网站被标注的标签 |
|---|---|
| http://www.google.com | google、searchengine、engine、web、search |
| http://www.technorati.com | blog、blogs、web2.0、news、search |
| http://www.indeed.com | jobs、jobsearch、career、job、search |
| http://www.krugle.com | programming、code、development、opensource、search |
| http://www.isohunt.com | torrent、torrents、bittorrent、p2p、search |
| http://www.koders.com | Programming、code、source、opensource、search |

表3.1列出的网站是检索“search”标签所返回的部分结果。由于标签检索返回的资源与标签具有一定的相关性，在本文的假设前提下，可以得出：用于查询的标签与返回资源的标签集之间也存在着语义相关。

### 3.2 扩展标签的选取标准

利用上节中相似标签的选取方法，可以得到原始标签 $t_0$ 的相关标签集合 $T=\{t_1,t_2,\cdots,t_{|T|}\}$，其中每一个 $t_i\ (1\le i\le|T|)$ 称为 $t_0$ 扩展标签。

对于查询扩展来说，初始查询 $Q=\{q_1,q_2,\cdots,q_{|Q|}\}$ 中含有多个词项 $q_j\ (1\le j\le|Q|)$，每个词项 $q_j$ 可以根据相似标签的选取方法获得相关的标签集合。依据返回的相关标签集合对每个词项 $q_j$ 进行语义扩展，这样就实现了对初始查询 $Q$ 的语义扩展。

从理论上来说，返回的相关标签集合中任何一个标签都有可能成为扩展词。本文给出一种扩展标签的选取评分标准，扩展标签的评分表示与原始查询的关联程度，这样我们就可以将评分高的前 $k$ 个扩展标签作为原始查询的扩展词。下面的重点就是如何构造扩展标签的评分函数。

Xu和Croft的研究[4]表明,利用词项之间的共现度来选取扩展词，能够取得更好的扩展效果。共现度是指两个词项在一定的文本窗口下共同出现的次数,本文定义的窗口是通过标签查询返回的每一个资源 $p$。由表3.1可以看出，每个资源都包含一定数量的标签，这些标签能够反应资源的内容和主题，因此可将这些标签当作资源的特征词项。这样，资源可以理解成为文档，标注资源的标签表示文档的特征词。一个查询 $q_j$ 可以获得多个返回资源，这些资源的集合称为局部资

源集$S_j$。下面定义扩展标签$t_i$和查询$q_j$在局部资源集$S_j$中的共现度$co_degree(t_i,q_j|S_j)$：

$$co\_degree(t_i,q_j|S_j)=\frac{\sum_{p\in S_j}\log(tf(t_i,p)+1.0)\times\log(tf(q_j,p)+1.0)}{\log(n)} \tag{1}$$

其中，$n$表示局部资源集$S_j$的资源总数，$tf(*,p)$表示一个词项在资源$p$中的出现次数。

对于原始查询$Q$的每个子查询$q_j$，可以根据$co_degree(t_i,q_j|S_j)$从$S_j$中选取与$q_j$共现度高的若干个词作为扩展词。但是，每个子查询$q_j$只能反应出原始查询$Q$的部分查询信息，简单的对各个$q_j$进行独立的扩展是不合适的。

针对这个问题，我们需要考虑每个扩展词$t_i$与整个查询$Q$的相关程度。假设原始查询$Q$中的每个查询词$q_j$之间相互独立，则可以根据$t_i$与每个$q_j$在$S_j$中的共现度来度量扩展词$t_i$与原始查询$Q$的相关度。实际上，查询$Q$中的各个子查询词$q_j$之间具有不同的重要性，有的子查询词显得比其他子查询词更为重要。根据这一特点，我们引入倒转文档频率（idf）来表示一个词项在全局文档集的重要性。因此，我们得出最终的扩展标签评分函数$Score(t_i,Q|C)$：

$$Score(t_i,Q|C)=\sum_{q_j\in Q}idf(q_j,C)idf(t_i,C)\log\left(co\_degree(t_i,q_j|S_j)+1.0\right) \tag{2}$$

其中，$idf(*,C)$定义为：

$$idf(*,C)=\log\frac{N}{df(*,C)+1.0} \tag{3}$$

全局资源集$C$表示为原始查询$Q$中每个子查询$q_j$返回的局部资源集$S_j$的并集。$df(*,C)$表示全局资源集$C$中出现某个词项的文档数目。

### 3.3 扩展标签的权重分配

查询扩展的另一个问题是如何对选取的扩展词进行权重分配。王斌和丁国栋等[12]提出了将扩展词的评分加入到传统的Rocchio公式中，并取得了很好的效果。因此，我们考虑利用这种扩展评分值来改善原有的权重计算公式。具体的权重分配函数$Weight(q'|Q')$：

$$Weight(q'|Q')=\alpha\cdot Weight(q'|Q)+\beta\cdot\frac{Score(q')}{MaxScore} \tag{4}$$

其中，$Score(q')$表示扩展词$q'$的评分值，可由$Score(t_i,Q|C)$公式计算得到；$MaxScore$为所有扩展词的最大评分值；$\alpha$和$\beta$分别设置为0.8和0.2。

## 4 实验设计

实验使用的是TREC2008相关反馈的测试数据集。数据大小约为424GB，包含25204669篇文档，共有264个Topics，这些Topics取自不同领域，长度在1-8个词项之间。针对264个Topics中的不同词项，我们从del.icio.us网站爬取了大约20万条相关的网页信息，其中包含60多万个候选扩展标签。对于一个标签查询，del.icio.us网站会返回一定数量的相关网页信息，我们提取其中前2000条资源信息作为一个子查询局部资源集，即$n$=2000。

用户标注行为的随意性导致了部分标签的不规范，因此在使用标签之前，需要对其进行预处理。本文主要处理以下两种不规范标签：（1）有些标签是用来反应个人需求，对资源没有描述作用，例如，“toread”、“2read”、“@read”。（2）有些复合标签是由两个具有独立意义的标签组合而成的，需将这部分标签分割成多个标签进行处理。例如，“java/programming”、“iraq_war”。

实验目的是为了说明根据本文提出的方法所选取的扩展标签与原始查询所表征的主题具有

较高的语义相关性。在查询扩展实验中，如果扩展词与原始查询的主题相关性不高，检索结果会低于未进行查询扩展的原始检索结果。在查询扩展的实验中，本文采用MAP值作为主要评测指标，并以P@10和P@20作为辅助的评测指标。

# 5 实验设计

## 5.1 扩展标签的选取结果

根据3.2节提到的相关标签选取方法，我们对264个Topics进行基于标签的语义扩展，并使用3.3中的权重分配公式为扩展词进行权重的分配。表5.1给出了“total knee replacement surgery”（Topic 812）和“imported fire ants”（Topic 820）两个Topics的扩展结果，这里选择了前14个权重最高的扩展词。

表5.1 前14个最相关的扩展标签表

| Topic 812 | Weight | Topic 820 | Weight |
|---|---|---|---|
| knee | 0.4 | ant | 0.4667 |
| surgery | 0.3316 | fire | 0.3897 |
| replacement | 0.3091 | import | 0.3688 |
| total | 0.2773 | science | 0.0392 |
| health | 0.0747 | news | 0.0315 |
| cosmetic | 0.0417 | biology | 0.028 |
| plastic | 0.0377 | pest | 0.0274 |
| software | 0.0363 | insect | 0.025 |
| windows | 0.0341 | game | 0.0233 |
| exercise | 0.0301 | home | 0.0218 |
| medical | 0.0298 | reference | 0.0215 |
| eye | 0.0273 | art | 0.0209 |
| injury | 0.0264 | animal | 0.019 |
| pain | 0.0234 | nature | 0.0188 |

以表5.1中“imported fire ants”（Topic 820）为例，扩展标签权重最高的3个词“ant”、“fire”、“import”就是原始查询中的词项，这说明权重分配公式更加偏重于原始查询的主题语义。其他权重较高的扩展词中，“science”、“biology”、“pest”、“insect”都与原始查询的主题具有很高语义相关性。但是，“news”、“game”等词与原始查询的主题就缺乏相关性，出现这种情况的原因是：在选取候选标签时，本方法是以单一词项为独立单位进行语义扩展的，这就会引入一些不相关的“噪音”标签。关于如何减少这种“噪音”标签的产生，这是我们将来的研究工作之一。

## 5.2 基于标签的扩展对检索性能的影响

表5.2给出的是，在TREC2008相关反馈数据集上不进行查询扩展与采用基于标签的查询扩展方法后的检索性能。其中，“无扩展”表示的是不进行查询扩展，直接使用原始查询词进行查询的检索性能，另一行“基于标签的扩展”就是采用本文提到的方法进行查询扩展后的检索性能。

表 5.2 无扩展和采用基于标签的扩展方法的检索性能

| 查询扩展方法 | MAP | P@10 | P@20 |
|---|---|---|---|
| 无扩展 | 0.2497 | 0.5143 | 0.4857 |
| 基于标签的扩展 | 0.2914（+16.7%） | 0.5347（+3.97%） | 0.5306（+9.24%） |

由上表可以看出，与未进行扩展的结果相比，基于标签的扩展方法在MAP、P@10、P@20等指标上都有一定提高，其中主要评测指标MAP值提高幅度较大，大约提高16.7%。对于查询扩展来说，如果扩展词与原始查询的相关性不高，会导致检索效果低于未进行查询扩展的原始检索结果。本文提出的基于标签共现的查询扩展方法比未进行扩展的检索效果有所提高，这表明该方法可以有效的提取语义相关的扩展标签，folksonomy的标签资源可以为查询扩展提供准确的扩展词来源。

## 6 结束语

本文通过挖掘 folksonomy 中标签之间的语义关联，采用基于标签共现的方法来获得语义相关的扩展标签。在利用词项与相关资源标签的共现度来衡量扩展标签的相关程度的基础上，该方法综合考虑了原始查询中词项之间的相互影响，使得选取出的扩展词与原始查询所表征的主题具有较高的相关性。实验表明，采用基于标签共现的扩展方法可以使检索的平均准确率提高 16.7%。

在本文的工作基础上，还有一些问题需要进一步分析研究：(1).原始查询中词项之间的依存关系问题。本文假设原始查询中各个词项之间是相互独立的。但是对于实际的查询来说，词项之间是存在着一定的依存关系的，利用这种依存关系，不仅可以挖掘出更多相关的语义信息，还可以减少“噪音”标签的产生。(2).挖掘 folksonomy 的社会化团体信息。本文主要是依据 folksonomy 中标签与资源的标注关系来获取语义关联的标签，并将其运用到查询扩展中，取得较好的效果。Folksonomy 中还包含有大量的用户信息，可以从中提取用户团体的一些属性信息，将这些信息与本文的方法相结合，实现针对特定用户团体的查询扩展。

## 参 考 文 献


[1] Qiu Y and Frei H. P. Concept based query expansion. In Proc. of SIGIR' 93, 1993, 160-169.

[2] Jing Y and Croft W. B. An association thesaurus for information retrieval. In Proc. of the Intelligent Multimedia Information Retrieval Systems (RIAO' 94), 1994, 146-160.

[3] G. Cao, J. Nie, J. Gao and S. Robertson. Selecting Good Expansion Terms for Pseudo-Relevance Feedback. In Proc. of SIGIR' 08, 2008, 243-250.

[4] Xu J. X and Croft W. B. Improving the effectiveness of information retrieval with local context analysis. ACM Transactions on Information Systems, 2000, 18(1):79-112.

[5] G. S. Atomiq. Folksonomy: social classification. http://atomiq.org/archives/2004/08/folksonomy social classification.html, August 2004.

[6] Laura Gordon-Murnane. Social Bookmarking, Folksonomies, and Web 2.0 Tools. Searcher, 2006, 14(6):26-38.

[7] Cameron Marlow, Mor Naaman, Danah Boyd and Marc Davis. Position Paper, Tagging, Taxonomy, Flickr, Article, ToRead. In Proc. of the 17th Conference on Hypertext and Hypermedia. Odense, Denmark, 2006.ACM Press New York: 31-40.

[8] S. Xu, S. Bao, B. Fei, Z. Su, and Y. Yu. Exploring Folksonomy for Personalized Search. In Proc. of SIGIR' 08, 2008, 155-162.

[9] H.S. Al-Khalifa and H. C. Davis. Exploring the value of folksonomies for creating semantic metadata. IJSWIS, 3(1), 2007, 13-39.

[10] P. Mika. Ontologies are us: A unified model of social networks and semantics. In Proc. of ISWC'05, 2005, 522-536.

[11] X. Wu, L. Zhang, Y. Yu. Exploring social annotations for the semantic web. In Proc. of WWW'06, 2006, 417-426.

[12] 丁国栋,白硕,王斌. 一种基于局部共现的查询扩展方法. 中文信息学报,2006, 20(3): 84-91.