硕 士 学 位 论 文

基于社会化标注的个性化检索方法研究

Personalized Search Based on Social Annotation

作 者 姓 名：管毅舟

学科、 专业：计算机应用技术

学　　号：21209213

指 导 教 师：林鸿飞 教授

完 成 日 期：2015 年 4 月 30 日

大连理工大学

Dalian University of Technology

大连理工大学学位论文独创性声明

作者郑重声明：所呈交的学位论文，是本人在导师的指导下进行研究工作所取得的成果。尽我所知，除文中已经注明引用内容和致谢的地方外，本论文不包含其他个人或集体已经发表的研究成果，也不包含其他已申请学位或其他用途使用过的成果。与我一同工作的同志对本研究所做的贡献均已在论文中做了明确的说明并表示了谢意。

若有不实之处，本人愿意承担相关法律责任。

学位论文题目：基于社会化标注的个性化检索方法研究

作 者 签 名 ：_____ 日期：____年____月____日

摘　要

随着搜索引擎的发展，用户对于信息检索的需求也日益提高。为了更好的满足用户对于查询的需求，个性化检索技术应运而生。其主要是用于解决在信息检索时如何根据用户兴趣返回对应的搜索结果的问题。在这一过程中，越来越多的资源被引入其中，用来构建用户的兴趣。随着 Web2.0 的发展，社会化标注系统得到了很好的普及，而社会化标注作为用户直接给出的资源，对应个性化检索有着很大的意义。目前，基于社会化标注的个性化检索的有效性已经得到了很好的验证，但依然存在一定的上升空间。本文的主要目的就是在于如何能够使用社会化标注更好地提升个性化检索的效果。

在使用社会化标注进行个性化检索时，要注意标签只是用户兴趣的一个方面，不能完全表示。同时不同用户给出的标签在质量上也有区别，这种质量也会对检索的准确性产生影响，同时标注的稀疏性问题也会给检索带来一定的困难。基于上述分析，本文的主要贡献包括以下两方面：

首先，仅使用社会化标注时，本文提出了基于用户相似网络的个性化检索方法。用类比 VSM 模型的方法计算标签对于用户的权重，并引入了用户在文档上的相似性这一概念，使用户相似度计算更加准确。同时将用户之间的相似关系看作网络关系，参考网络中节点互相影响的方法在降低检索复杂度的同时解决标注稀疏的问题，从而使检索结果对于用户而言更加可靠。实验证明该方法可以在减小消耗的同时一定程度提升检索质量。

其次，为了更好地构筑用户的兴趣，本文提出了一种融合用户相似度和用户质量的个性化检索方法，该方法在使用社会化标注的基础上引入了网页分类信息共同计算用户之间的相似度，并且通过两种资源结合改进的 Social PageRank 算法对于用户的质量也进行判断，从而在扩展时提高质量，进而提升检索效果。本文实验证明该方法对于个性化检索的效果有较大提升。

关键词：个性化检索；社会化标注；用户相似度

The Format Criterion of Master's Degree Paper of DUT

Abstract

As the development of search engine, users have more requirement for information retrieval. For better satisfied it, personalized search method is invented. The main propose of these methods are solve the problem that how to give results that based on user's interest when in retrieval. In this process, more and more resource is add to it to build users' interest. Because of the development of Web 2.0, social annotation system is well popularize. Social annotations are the resource which come from users directly, so it's very useful for personalized search. By the time, using social annotations in personalized search is tested and is proved to be effective. But there are still some way to advance it. The main purpose of this paper is to explore how to use social annotations to raise the effect of personalized search.

There are something which should be pay attention to when use social annotations. One of them is that annotations are one side of users' interest, which can't completely represent it. And the annotations from different users may have different quality, which can influence the result. Besides, the sparsity of annotations is also difficult to solve. Depend on these reason, the main contribution of this paper are:

First, when only use social annotations, we promote a personalized search method based on the network of users' similarity. We use a method which is similar to VSM to calculate the weight of users' tags, and we add the similarity of users on the same document to it. Then we consider the relation of similar users as a network, and use the method like the interaction of nodes to solve the sparsity problem and reduce complexity of search. So that the result will be more reliable. The experiment prove that the method can improve the effect and reduce the consumption.

Then for better constructing user interest, we promote a personalized search method reconcile with user similarity and user quality. We add web categories to the method that only use annotations to calculate users' similarity. And we also user these two resource in the modified Social PageRank to judge users' quality, so that the extension of annotation and the quality of retrieval will be improved. The experiment shows that this method can obviously improve the result.

Key Words：Personalized Search; Social Annotation; User Similarity

目　　录

1 绪论

1.1 研究背景

近年来，Web2.0 的应用在社会上越来越普及，比如社会化标注系统、电子商务系统以及个性化网络，这些都不断渗入我们生活的方方面面。而我们所处的网络，也正朝着个性化网络迈进。在这样的个性化网络中，用户可以通过标注、评论等多种方式在在线网页上留下自己的观点。作为其中比较重要的一种，用户在网络中的标注行为经常对网页内容起到了很好的概括，并反映了用户对网页的想法和态度。这样的标注同时也是对网络中丰富资源的分类，并且这种分类不需要具有系统知识的专家进行维护，因此这种标注称为社会化标注（social annotation），该分类方法称为大众分类 (folksonomy)[1-3]，而提供这种标注功能的系统称为社会化标注系统(Social Annotating System)或大众分类系统(Folksonomy System)，例如， Flikr[4], Del.icio.us[5], Bibsonomy[6] 等都是网络中为人熟知的社会化标注系统。随着社会化标注系统的不断普及和日趋流行，其所具有的一些良好特性使其在信息检索领域中的应用也越来越广泛[7-10]。

与此同时，随着用户网络检索能力的上升及信息检索技术的成熟，用户对于信息检索服务准确性的要求也越来越高。然而，用户在通常情况下提交的查询都比较简短精炼，为其查询目的的高度概括，并不能完全准确地表达真正的需求。同时，每个用户都有其独有的兴趣，其在检索的时候，往往与其兴趣相关。即使查询完全相同，也可能因为用户的兴趣不同而需要返回不同的结果。为了满足用户这种个性化的需求，个性化检索技术应运而生[11]，并且逐步发展。综合来看，个性化检索技术目前的关注点分别都眼于用户兴趣的构建上。最开始的研究普遍使用用户检索的历史作为兴趣的来源[12-13]，随后一些学者开始使用外部资源对用户兴趣进行挖掘。例如，网页中的本体资源[14-15]、网页分类信息[16-17]以及用户的眼动轨迹[18]等。而在用户兴趣使用的方法上面，又分为了查询扩展[19]和重排序[20]两种方法。

2008 年，Xu 等人[21]的研究指出并验证了基于用户社会化标注构建的用户兴趣能够有效提高个性化检索的效果，为使用外部资源的个性化检索技术指出了一个新的方向。然而，在真实的网络中，每个用户能够浏览和标注的网页总是有限的，并且该数量对比于网页总数总是极小的，因此社会化标注总是稀疏的。为了解决这个问题，Xu 等人[22]提出了使用社会化标注计算用户相似度并进而扩展标注信息的方法，有效解决了标注的稀疏性问题。但由于存在一些诸如“fun”等网页通用的标注、并且用户给出标注时没有约定的规则，往往带有一定的随意性，所以计算用户相似度时不能简简单单考虑标签

的出现次数。因此，在基于社会化标注的个性化检索过程中研究如何能够更准确地计算用户相似度并进行用户标注的扩展。这也是本文即将要进一步研究的内容。

1.2 研究现状

个性化检索是近年来被广泛关注的一个研究领域，其主要目的是解决返回给用户的检索结果无法满足用户个性化的需求这一问题。该技术利用计算机信息学等技术和方法，通过一定的算法将用户兴趣和偏好的因素加入到网页排序模型中，使返回的结果根据兴趣不同而有所差异，以满足用户的个性化检索需求，进一步提高检索的用户满意度。

在信息检索技术不断成熟的现在，个性化检索技术也已经得到了很大的发展，出现了多种个性化检索的方法。个性化检索中大多数的方法都是基于用户兴趣构建的，依据使用的资源不同，个性化检索的方法也分为不同种类，其中比较常用的有基于用户检索历史和基于网页相关的外部资源的两种。

基于用户检索历史的个性化检索方法所使用的资源也是多种，其中为人熟知的有用户查询历史[23]、浏览历史[24]、用户检索的任务[25]、用户检索的目的[26]，甚至是用户浏览检索结果的眼动轨迹[18]。对上述数据进行解析，获取用户所频繁关注的词语，并基于这些词语构建用户的兴趣和偏好。之后将兴趣的表示加入检索模型中，对查询进行扩展[19]或者对返回的文档集合进行重新排序[20]，以达到个性化搜索结果的目的。在这方面比较广为接受的有用户点击日志，其可用性因为已经被验证[18]，所以其被用来挖掘用户隐藏兴趣并与话题敏感的 PageRank 组合进行个性化[27]。

近年来，越来越多的研究开始着眼于使用新的外部资源进行个性化检索，这些资源被作为用户兴趣和偏好的来源，在资源上进行挖掘可以得到用户潜在的信息，这既是使用基于网页相关的外部资源的个性化检索技术。在个性化检索领域常用的外部资源一般有网页本体资源[14-15]和网页分类信息[16-17]等。随着网络的进一步发展及 Web2.0 的进一步普及，社会化标注的应用越来越受到关注，在研究方面也取得了一定的进展。其中比较为人所熟知的有语义检索[28-29] 、网页检索[30-31] 、标签推荐[32-33] 、资源分类[34-35]及查询扩展[36-37]等。

在社会化标注在个性化检索的应用研究中，Xu 等人提出了一种基于标注系统中用户和网页属性的相似度对结果进行个性化的方法[21]。Bouadjenek 基于此进行了扩展，其引入了个性化匹配分数以解决网页匹配的问题[38]。然后 Noll 使用标注的频率作为权重并将所有网页中标注的权重标准化以赋予用户属性更大的作用[39]。为了使用户属性的各项权重更加合理，Vallet 使用了 BM25 模型[40]代替了原来的 VSM 模型[41]。但受限于用户

本身的局限性，社会化标注往往比较稀疏，所以这些方法的效果也受到了限制。为了解决这一问题，Xu 等人提出了一种叫做双重个性化排序方法的算法（Dual Personalized Ranking，后文简称 *D-PR*）[22]，通过给出的标注计算用户之间的相似度，并通过每一用户的相似用户扩展对应用户的标注，使每一个用户的标注都覆盖到更多的网页。之后基于扩展后的标注计算用户属性和网页属性，进一步计算出个性化匹配得分与网页查询相似度进行插值，最后根据最终得分进行网页排序。

虽然上述方法在个性化检索领域取得了不错的效果，但其在判断相似用户时，使用社会化标注的方式比较简单直接，所以用来进行扩展的相似用户的正确性难以得到保证，从而使个性化检索的准确率还存在进一步提升的可能性。

基于以上的研究基础，本文将社会化标注信息作为个性化检索的基本资源，探索使用社会化标注判断相似用户并进行标注扩展，利用扩展后的标注计算文档最终排序得分并进行重排序，旨在改进个性化检索的效果，满足用户日益增长的个性化信息需求。

1.3 研究动机及意义

在网络资源逐渐丰富的今天，搜索殷勤在人们的生活中扮演的角色也越来越重要，随时随地满足人们对信息的需求。在实际使用中，不同用户在检索时往往带有自己特定的目的和目标，这使千篇一律的搜索结果已经无法让用户感到满意，所以检索结果的个性化日益重要。这就促使了个性化检索技术进入人们的视野，根据用户的不同需求来返回“定制”的结果成为了热门话题。在这大环境下，基于查询历史数据和基于与网页相关的数据的个性化检索方法取得了一定的成果，其中基于社会化标注的个性化检索方法在这其中逐渐发展了起来。

需要注意的是，在社会化标注系统中，用户的标注行为往往具有随意性，这就造成了标注在网络中分布的不规则性和随意性，也可能会包含一些无意义的、甚至是垃圾的信息。这些因素导致在个性化检索中判断相似用户时的准确性降低，从而使用户兴趣和偏好的构建准确性降低，进而影响个性化检索结果的效果。为了改善这一问题，有必要提出一种能够有效使用社会化标注寻找相似用户进行用户兴趣和偏好构建的方法，以此提升最终搜索结果的满意度。

在个性化检索中判断相似用户时，用户标注行为的质量、数量及标注本身的意义都能对用户的相似度产生影响，从而使用户兴趣和偏好的构建有所不同。通过挖掘用户标注行为，具体指标注对于用户兴趣的意义及不同用户间在标注上的特性，能够更加准确地筛选出目标用户的相似用户，从而更好地建立目标用户的兴趣和偏好，提升检索结果的准确性。

1.4 论文组织结构

本文的内容组织如下：

第一章：主要介绍本文的研究背景、研究现状，并指出了本文研究的动机及意义；

第二章：主要介绍了本文研究过程中涉及的一些相关技术及其实现方法，同时介绍了社会化标注、网页分类信息的概念、表示方法，以及目前在研究中被普遍使用的算法和技术等；

第三章：介绍本文所提出的基于用户相似网络的个性化检索方法。该方法将同一用户的社会化标注看做一个集合，类比于一篇文档，并将在信息检索时对文档进行处理的方法引入，以消除共性、突出特性，从而使相似用户的判断更为准确，同时将用户与用户之间的相似关系看成网状结构，通过用户节点之间互相影响的方法解决稀疏性的问题，从而提升搜索结果的准确性；

第四章：介绍了融合用户相似度和用户质量的个性化检索方法，首先分析了与用户兴趣相关的标注行为及网页浏览行为，提出了一种结合两者计算用户相似度的方法，并且对用户的质量也进行计算。并在此基础上筛选出相似用户融合用户质量进行社会化标注的扩展，进而提升个性化检索的准确率。

结论：总结全文，并对今后的研究工作进行展望。

2 相关理论知识、技术及方法

2.1 相关技术

2.1.1 信息检索技术

信息检索技术发展自 19 世纪末，最早来源于在图书馆中对文献进行查找。随着计算机科学的发展，现在信息检索指信息按一定的方式组织起来，并根据信息用户的需要找出有关的信息的过程和技术。在信息检索技术发展的过程中，形成了多种各有特点的检索模型。

向量空间模型(Vector Space Model, VSM) 是信息检索当中比较常用的一种模型，这种模型主要是将需要衡量的文档或者查询都映射至项空间一个向量空间中(如图 2.1)，其中的项通常为词语。即是将需要文档或查询经过一定的处理转换成同一向量空间中的向量进行表达，然后通过衡量其在同一向量空间中向量的距离，获得这些向量所代表的资源的相似度度量，最终依据相似度从高到低对文档集中的文档进行排序，获得该查询下的文档排序结果[42]。在个性化检索领域，VSM 适用的领域又扩大了，其向量空间不光可以指文档中的词，也可以指用户标注过的标签以及网页所属的类别。当处理网页或查询相似度时向量空间时使用词语，处理社会化标注时时使用标签，而处理网页分类信息时则使用网页类别。

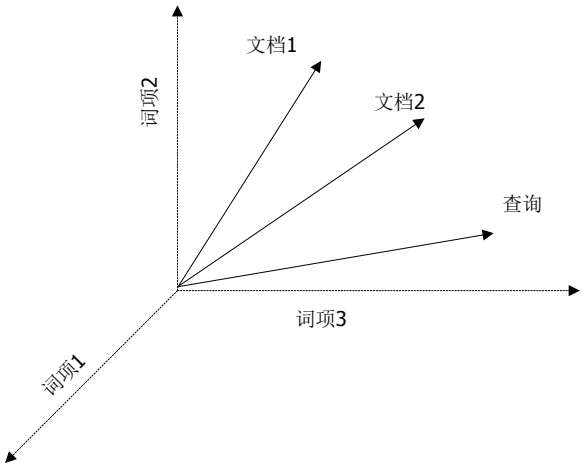

图 2.1 空间向量模型中文档与查询表示

Fig. 2.1 The representation of documents and queries in VSM

以最传统的 VSM 模型为例，假设词语向量空间为 n 维，则文档 *D* 在其上的投影可以表示为可表示为：

$$D_i = (d_{i1}, d_{i2}, \cdots, d_{in}) \tag{2.1}$$

查询 Q 包在向量空间上的投影可表示为：

$$Q_i = (q_{i1}, q_{i2}, \cdots, q_{in}) \tag{2.2}$$

在此基础模型上，对权重的计算衍生出了多种方法，其中在信息检索中比较常用效果也较好的一种是 TF-IDF 算法：

$$tfidf_{i,j} = tf_{ij} \times idf_i \tag{2.3}$$

$$tf_{i,j} = \frac{n_{ij}}{\sum_k n_{k,j}} \tag{2.4}$$

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|} \tag{2.5}$$

其中，$tfidf_{i,j}$表示词语t_i在 TF-IDF 算法下得到的值，即其在文档或查询向量中词应有的权重；$tf_{i,j}$表示词语t_i在文档d_j中出现的频率，这个数字是对词数的归一化，以防止它偏向长的文件；idf_i表示词语t_i在整个文档集中的倒排文档频率，可由式(2.5)得到，是一个词语普遍重要性的度量，其中$|D|$为语料库中的文件总数，$|\{j : t_i \in d_j\}|$为包含词语t_i的文件数目。该方法的基本思想是：如果某个词语在一篇文档中出现的频率高，并且在其他文档中很少出现，则认为此词语具有很好的类别区分能力。这种权重估计对那些在多数文档中都出现的，对区分文档意义不大的词语进行一定的惩罚，从而降低了文档间的共性，突出了文档的特性，进而能有效降低不同类型的文档相似的可能性。

之后计算文档和查询的相似度就转化为计算两个向量在向量空间中的距离，在这一方面比较好用的公式为余弦相似度。即给定两个 n 维向量 $A=(A_1,\ldots,A_n)$和 $B=(B_1,\cdots,B_n)$，余弦相似度 $Sim(A,B)$的计算方法如下所示：

$$Sim(A,B) = Cos(A,B) = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2.6}$$

向量空间模型的建立有一个主要假设：文档与查询中的词语之间相互独立、互不相关。这一假设使得文档和查询的每一个词都可以用于其他词语不相关的方式来表达。但是事实上文档和查询中的词语与词语间都会存在一定的语义或意义上的关系，这在文档中表现得更为明显。向量空间模型未考虑了这种关系，因此使用该模型的检索效果不可能非常精确。而模型中的独立性假设也使要在其中引入词语间的关系进行改善变得非常困难。

而对于文档的信息检索而言，通常最简单的想法就是考察词语在文档中是否出现，出现即代表相关，不出现则代表不相关。这样的相关规则叫做二元相关，即分为“出现”和“不出现”。基于这个基本认知以及和 VSM 一样的词语独立性假设，BM25 模型被提了出来[40]。该基本想法跟 VSM 类似，将查询分解为词语，对于一篇带检索文档 d，计算每个词语与其的相关性并求和，得到文档得分后进行排序。其得分计算公式可以表示为：

$$Score(Q,d)=\sum_{i}^{n} W_i \cdot R(q_i,d) \tag{2.7}$$

其中 W_i 表示查询中词语 q_i 的权重，$R(q_i,d)$ 表示词语 q_i 与文档 d 的相关性得分。根据公式,最重要的一点就是权重 W_i 的确定及词语与文档相关性得分的计算。

在权重的计算方面，包括上述 TF-IDF 方法在内，有多种计算权重的公式，但应用最多也最广泛的为 IDF 方法。其仅保留了 TF-IDF 中与词语出现的文档数相关的一部分，保证了同一词在不同文档间拥有相同的权重，使区分度高的词语对于查询和文档的相似性得分的计算有更大的贡献。

在计算词语与文档相关性得分方面，BM25 模型的一般公式为：

$$R(q_i,d)=\frac{f_i\cdot(k_1+1)}{f_i+K}\cdot\frac{qf_i\cdot(k_2+1)}{qf_2+k_2} \tag{2.8}$$

$$K=k_1\cdot\left(1-b+b\cdot\frac{dl}{avgdl}\right) \tag{2.9}$$

其中，k_1、k_2、b 为调节因子，通常根据经验设定取值大小，一般 k_1=2、b=0.75；f_i 为 q_i 在 d 中出现的频率，qf_i 为 q_i 在查询中出现的频率。dl 为文档 d 的长度，$avgdl$ 为所有文档的平均长度。由于绝大多数情况下，q_i 在查询中仅出现一次，所以上式可以简化为：

$$R(q_i, d) = \frac{f_i \cdot (k_1 + 1)}{f_i + K} \quad (2.10)$$

从 K 的定义中可以看到，参数 b 的作用是调整文档长度对相关性大小的影响。b 越大，文档长度的对相关性得分的影响越大，反之越小。而文档的相对长度越长，K 值将越大，则相关性得分会越小。这可以理解为：当文档较长时，包含 q_i 的机会越大，因此，同等 f_i 的情况下，长文档与 q_i 的相关性应该比短文档与 q_i 的相关性弱。

2.1.2 个性化检索技术

近几年来，个性化检索技术在用户需求的促进下不断发展，并取得了很多成果。在这段发展历程中，诞生了各种各样的个性化检索方法，而其中大部分方法的思想都是基于用户兴趣和偏好的挖掘构建。而构建的方法又分为基于查询历史及基于外部资源的方法。

基于查询历史的个性化检索技术，其主要是与用户检索记录相关的各项资源中挖掘能代表用户兴趣和偏好的信息，其中大部分仅仅从查询日志或记录中提取，也有一些研究借助额外手段收集一些平常不易获取的信息。

从查询日志中挖掘用户兴趣和偏好的方法通过对用户过往检索的历史记录当中寻找与其兴趣与偏好相关的信息，从而构造用户在检索时候所关注的焦点，在检索时在检索模型中融合进用户的兴趣，使结果偏向用户感兴趣的方向，从而达到个性化检索结果的目的。这种方法方便而有效，最大的优点是不需要服务器额外的话费，因为其所需要的资源都来源于服务器日常的记录，是搜索引擎的自然产物，使用起来代价较小。使用这一资源的个性化检索方法包括基于用户历史浏览文档的[43]、基于用户历史查询的查询词的[44]以及基于其余用户查询历史的[45]。然而，这种个性化检索方法存在一定的问题，主要是用户的查询历史很有可能包含用户个人的隐私在内，所以无法完全确保在不涉及隐私的情况下进行个性化检索，这对于其在真实环境中的应用是一个很大的阻碍。

借助额外手段的方法则是在用户过往查询时或开始查询前，利用与查询本身不相关的方法收集一些额外的信息以获得用户的兴趣和偏好。这种方法的关键是寻找与用户兴趣相关，又是用户会愿意主动提供的信息资源。在这一方面比较常用的是当前网络中广泛存在的调查问卷的形式，其通过在用户检索前让用户主动提供些有用信息，以达到直接获取用户兴趣的目的。另外也有使用眼动轨迹的研究方法[18]，使用眼动仪记录用户检索后浏览检索结果的情况，获得用户高度关注的结果，并以此进行挖掘构造兴趣表达，之后融合进检索模型。

随着用户对个性化的需求越来越明显以及自我保护意识的增强，基于查询历史的方法越来越表现出局限性，用户的隐私往往是这类方法无法解决的一大难题。为了寻求突破，更多的研究开始关注与文档相关的外部资源，这就形成了基于外部资源的个性化检索技术。目前常用的外部资源一般包括文档中的语义本体[14]、网页分类信息[16]以及用户的社会化网络[46]等。

基于外部资源的个性化检索技术利用已存在的与网页或与用户相关的信息，将与用户相关度高的信息加入用户兴趣构成，从而实现搜过结果的个性化。这种方法由于所用资源仍然是与用户间接相关，所以效果往往很难达到一个很理想的水平。

随着社会化网络的兴起，由用户直接主动给出的社会化标注越来越引起研究者的关注。社会化标注拥有良好的特性：一方面，其由用户直接给出，是与用户直接相关的信息；另一方面，其在网络中相对叫易于获取，且很少包含用户的隐私，所以使用的时候相对较为方便。因此，许多信息检索方面的研究都围绕社会化标注展开。

Xu 等人[21]提出了一种基于用户社会化标注与文档相似度的个性化检索方法，将用户主动给出的社会化标注看作用户兴趣的直接表达，一个用户给出的全部标注和一篇文档得到的全部标注被看作其各自的用户属性及文档属性，之后在计算传统的查询与文档的相似度之外，加上用户和文档属性的相似度，最终达到个性化的目的。之后为了将其扩展到没有文档的网络视频等资源上，使方法能在更广阔的领域得以应用，Bouadjenek 等人在原来的检索模型中加入了查询与文档属性的相似度，以使非社会化部分的得分更加准确[38]。

上述研究取得了一定的效果，但依然受到了一定的局限，这主要是由于网络中文档的数量相当庞大，而每个用户能浏览的只是其中一小部分，这就造成了社会化标注具有很大的稀疏性。为了解决这个问题，Xu 等人[22]提出了基于社会化标注计算用户相似度并扩展标注的个性化检索方法。其通过用户属性相似度判断出相似用户，之后使用相似用户的标注对目标用户进行扩展，最后使用扩展后的用户标注计算文档得分进而得到个性化检索结果。本文在这个方法的基础上，对其判断相似用户的方法进行改进，进而改进个性化检索结果的效果。

2.2 相关知识

2.3.1 社会化标注

随着社会化网络的兴起，网络资源日趋丰富，社会化标注系统越来越受到关注。在真实的网络环境中，社会化标注不仅仅为网页文档服务，同时其还可以附加在音乐、视频等多种形式的内容上。在这一方面，最为著名的当属 Flikr 和 Del.icio.us，前者提供了

对于图片的标注及检索功能，而后者则让用户可以对各种网站进行收录和标注。另外还有一种也广为人知，那就是视频网站，用户在网站中可以浏览感兴趣的视频，并对其留下评论、收藏或分享。这些社会化标注平台已经深入当今人们的生活之中，在其上人们可以根据自己的喜好对于海量的在线资源表达自己的看法。在用户的共同作用下，网络便拥有了充足的社会化标注，由于这些社会化标注是由用户按照自己的想法和观点自由为资源标注的，所以这些标注中包含了丰富的与用户个人相关的信息。而这些标注无形中也形成了对于网络资源的概括，这些概括也形成了一种对于资源的分类，这称为大众分类法(folksonomy)[4-5]。

表 2.1　社会化标注示例

Tab. 2.1　Example of social annotations

用户名	文档 d_1 的标注	文档 d_2 的标注	文档 d_3 的标注	文档 d_4 的标注	文档 d_5 的标注
Alice	English Comedy Interesting	Boring	Chinese Comedy Interesting	English	(null)
Bob	(null)	Action Interesting	Boring	Interesting	(null)
Carl	Comedy Interesting	(null)	Interesting	Boring	(null)
David	(null)	(null)	Chinese Comedy Interesting	Action Boring	Chinese Interesting

社会化标注由于来源于用户，所以其中包含着丰富的与用户相关的信息可供挖掘。用户的标注可以反映出其对于网络资源的喜好或态度，以及对于其中内容的思考和评价。同时用户和用户、用户和网页之间也因为社会化标注而产生了丰富的联系[47]。例如，表 2.1 中列出了部分用户对于与电影相关的部分文档的社会化标注。从表中可以看出：：1、*Alice*、*Carl* 和 *David* 偏好喜剧电影而不喜欢动作电影，而 *Bob* 喜欢动作电影而不喜欢喜剧电影。2、d_1、d_3 与喜剧电影相关，而 d_2、d_4 与动作电影相关。这就是社会化标注中待发掘的信息。正因为这其中丰富的语义信息，越来越多的研究着力于使用社会化标注，希望从中挖掘出有价值的内容。随着其被越来越广泛的采用，社会化标注所涉及的面也越来越广，其已渗入包含语义检索[28-29] 、网页检索[30-31] 、标签推荐[32-33] 、资源分类[34-35]及查询扩展[36-37]等多个方面。

由于本文主要关注的是网页文档的检索，所以使用的社会化标注信息全都来源于Del.icio.us，因此本文所介绍的标注的结构和特点，都以 Del.icio.us 上的标注为例。

在 Del.icio.us 中，用户可以对互联网中的网页进行收藏，并在其后附加标签。网站会将这些标注行为收集记录，形成如下形式的社会化标注：

<用户，标注网页，标注标签>

为了下文叙述方便，我们用 U 表示用户集合、D 表示文档集合，T 表示标签集合，则每一个标注信息就对应一个三元组 $(u, t, d) \in U \times T \times D$，代表用户 u 使用了标签 t 标注了网页 d。而标注信息集合就是 $U \times T \times D$ 的子集 $F(U、T、D)$。在表 2.1 所示例子中，用户集 U={*Alice,Bob,Carl,David*}，而网页集合 $D=\{d_1, d_2, d_3, d_4, d_5\}$，标注集合 T ={*English, Comedy,Interesting,Boring,Chinese,Action*}。对这样的记录进行挖掘，可以从中得到很多对信息检索有价值的内容。

社会化标注之所以能在信息检索中起到良好的作用，主要是因为其所具有的一些特性：

（1）关键词特性：由于用户通常给出的网页标注都只包含简单几个词，所以通常都为用户理解的高度概括、文档内容的精炼表达。所以标签可以看做是文档内容的关键信息，在文档中起着关键词的作用。因此将标注纳入信息检索领域考虑可以很好地改善检索效果。

（2）结构性：社会化标注由用户自由在网络中给出，所以用户是根据自己对网页的喜好程度选择性地标注网页，这种标注关系使大部分互联网用户都可以参与进来。而这海量的用户共同的标注行为也带来了海量的用户和网页的关系，这样的关系也事信息检索的有利资源。

（3）多样性：社会化标注来源于网络中不同的用户，而不同用户看网页的侧重点也不完全相同，这就造成了对统一资源的理解不同。所以一个网页上多个用户的标注就从不同侧面反映了网页内容，对这些内容进行总结，可以得到多个角度的概括和总结。

社会化标注的以上特点，使其能良好辅助信息检索，改善检索的准确性。但需要注意的是，在使用社会化标注时，其也存在一些固有的问题。首先，由于标注都有用户独立给出，所以其往往都遵循用户个人的语言，这就使在研究时对于标注信息的理解出现了困难。甚至有用户会根据自己的想法对词语进行组合，造成标注信息不是原有词语的现象。其次，用户在给出标注时都倾向于简单明了的词语，有不少标注词是其对网页喜好的表达，这就使标注中存在一些“泛用词”，比如“funny”、“Interesting”等，这些词语虽然也属于标注，但包含的有效信息很少，对于信息检索而言，价值意义不大。

最后，由于互联网中每个用户能浏览的网页总是有限的，所以其能标注的网页也是有限的，这就是标注的分布相对较散，出现了稀疏性的问题。如何降低稀疏性对社会化标注研究带来的局限性也就成了重点之一。

2.2.2 交互增强理论与算法

社会化标注系统主要包含资源、用户、标签这样三种类型的实体，这三种实体之间出了表面上的结构化关系上，还存在着相互增强的关系[50]，这样的关系可以通过下面两个规则进行描述：

（1）使用高质量标签标注高质量资源的用户称之为高质量用户；高质量用户使用高质量标签标注的资源称之为高质量资源；高质量用户标注高质量资源使用的标签称之为高质量标签。

（2）相似的标签通常被具有相似兴趣的用户分配给相似的网页资源。在社会化标注系统中，不同形态的相似标签能够通过这些用户标注的相同网页获得。

这种相互增强的关系称为交互增强理论(Mutual Reinforcement Principle)。交互增强理论显示出在社会化标注系统中用户、资源以及标签之间存在着更深层次的关系，其彼此之间正向相关。Social PageRank 算法正是基于这一原理提出的。

Social PageRank(SPR)算法[37]由 Bao 等人提出，该算法使用迭代的方法估计社会化标注系统中资源的质量。与其他基于交互增强理论算法不同的是，SPR 算法的基础在于构建资源到用户，用户到标签，标签到资源的转移矩阵，并根据三者间的两两关系将矩阵作为权重迭代计算三者的质量。假设有 N_T 个标签，N_R 个网页资源，N_U 个用户，令 M_{RU} 表示资源到用户的关联矩阵，为 N_R*N_U 维；令 M_{TR} 表示标签到资源的关联矩阵，为 N_T*N_R 维；令 M_{TU} 表示标签到用户的关联矩阵，为 N_T*N_U 维。关联矩阵中的每一个权重的初始值为相应两个实体之间关联的数量，比如 $M_{RU}(r_i,u_j)$的值为用户 u_j 标注资源 r_j 的标签数。另外，为了保证迭代的顺利进行，需要为每个资源定义一个初始的质量，本文利用 P_0 来表示。具体公式如（2.11）所示：

$$\begin{aligned} U_i &= M_{RU}^T \cdot R_i & T_i &= M_{UT}^T \cdot U_i \\ R'_i &= M_{TR}^T \cdot T_i & T'_i &= M_{TR} \cdot R'_i \\ U'_i &= M_{UT} \cdot T'_i & P_{i+1} &= M_{RU} \cdot U'_i \end{aligned} \qquad (2.11)$$

其中，R_i, U_i, T_i 分别表示在第 i 次迭代后资源、用户以及标签的质量向量，而 R_i', U_i', T_i'表示计算过程中三种实体的结果中间值。上述公式指出，用户的质量与资源的质量直接相关，可以通过资源质量乘以资源到用户的关联矩阵的转置得到。而标签质量也与用

户质量、资源质量与标签质量也存在类似的关系。也正因为如此，对该公式进行训话你迭代，就可以得到三种实体的向量收敛值，即表达了三种实体的质量。

算法的收敛性已经由 Bao 等人证明，同时他们也证明了该算法的有效性，为基于社会化标注的研究提供了一种很好的模式，并使标注、用户和资源变成了可以有效评价的实体，为后续其他研究提供了方便。

2.3 相关方法

2008 年，Xu 等人[21]提出了一种在指定用户 u 提出查询 q 时，计算文档 d 的得分的公式（*UP-PR*）。其认为公式应该包含两个部分：（1）基于文本匹配分数的部分 $Score(q,d)$，反映了文档 d 与查询 q 在文本统计上的相似程度；（2）基于属性的相似度 $Sim(p_u, p_d)$，反映了用户 u 对于文档 d 的兴趣，通过计算用户属性和文档属性的相似度得到。所以公式为：

$$R(d, q, u) = \alpha \cdot Sim(p_u, p_d) + (1 - \alpha) \cdot Score(q, d) \quad (2.12)$$

其中 p_u 为用户属性，代表用户对于网页的兴趣和偏好；为 p_d 文档属性，表示所有用户对于这篇文档的理解和观点。这两种属性都使用 VSM 模型，分别对于用户给出的标注和文档获得标注进行统计得到，向量的权值即为标签出现的次数。

用户在进行检索时，往往会碰到一些包含文本内容较少的资源，如视频、图片等，对于这类资源，其 $Score(q,d)$得分往往较低，所以导致排序得分也会较低。所以上述排序公式是偏向文本内容的，会给文档图片的混合检索带来问题。为了解决这一问题，Bouadjenek 等人[38]在 2013 年提出了一种社会化个性排序方法（*SoPRa*），对于 *UP-PR* 进行了扩展。其引入了一个新的非个性化的匹配分数：计算查询 q 与文档的社会化总结的相似度的匹配得分 $Sim(q, p_d)$。这一分数说明了查询与文档的社会化总结的相关程度，而引入的理由是因为社会化标注是网页资源的很好总结并且这能在排序时给包含较少文本内容的社会化资源带来更多的信息。因此，公式中与查询相关的分数变成了两个：

$$R(d, q, u) = \alpha \cdot Sim(p_u, p_d) + (1 - \alpha) \cdot [\beta \cdot Sim(q, p_d) + (1 - \beta) \cdot Score(q, d)] \quad (2.13)$$

然而在真实的网络环境中，不同用户的标注对于网页内容的贡献应该是不一样的，在查询时的价值也会是不一样的。同时由于标注具有的稀疏性，导致对于用户的衡量也是无法完全的，因为存在大量用户没有看过的网页。这两个问题共同限制了 *SoPRa* 的准确率。为了解决这个问题，Xu 等人[22]在 2014 年对其进行了改进，提出了一种双重个性化排序方法（*D-PR*）。该方法定义了两种新的属性：扩展用户属性及个性化文档属性，

区别对待不同用户的标注，以更好的获得用户的兴趣以及总结用户对于文档的个性化的观点。这是因为对比于用所有用户的标注来作为文档的概述，显然使用用户的个性化观点更为合适。对于用户 u，其扩展属性 p'_u 的定义为其所有个性化文档属性的总和。所以该方法的公式为：

$$R(d,q,u) = \alpha \cdot Sim(p'_u, p_{u,d}) + (1-\alpha) \cdot [\beta \cdot Sim(q, p_d) + (1-\beta) \cdot Score(q,d)] \quad (2.14)$$

公式中 $p_{u,d}$ 就代表了文档 d 对于用户 u 的个性化文档属性，根据方法，其计算公式为：

$$p'_u = \sum_{i=1}^{|D|} p_{u,d_i} \quad (2.15)$$

为了区别对待不同用户的标注，得到文档对于用户的个性化文档属性，该方法采用了基于用户属性的相似度作为加和系数。因为用户给出的标注都来源于用户的思考，所以其属性也就代表了他全部的想法，于是通过计算用户属性相似度可以很好区别用户应该被重视的程度。在计算出用户之间相似度之后，在每一篇文档上，将用户相似度乘以相应用户给出的标注并求和，就可以得到文档对于用户的个性化文档属性（自身的相似度相当于 1）。计算用户相似度及使用相似度计算个性化文档属性的公式如下所示：

$$PerSim(u', u) = Sim(p_{u'}, p_u) \quad (2.16)$$

$$p_{u,d} = \sum_{i=1}^{|U_d \cap U_T|} (v_{u_i,d} \cdot Persim(u_i, u)) \quad (2.17)$$

其中 U_d 为标注过文档 d 的用户，U_T 为用户 u 的相似用户。在上述公式中，VSM 模型依然被大量采用，这主要是因为其使用起来简单方便并且有效[42]。

2.4 本章小结

本章主要介绍了本文研究中所涉及的理论基础知识，包括信息检索相关概念以及两种常用且效果较好的检索模型，同时简要介绍了个性化检索技术的意义，并简单介绍了目前基于查询历史和基于外部资源的个性化检索技术。同时，本章还介绍了社会化标注的相关概念和特点等，并进一步介绍了交互增强理论和一系列基于社会化标注的个性化排序方法。本章介绍的内容主要是为后面章节中由本文提出的两种以社会化标注为基础，计算用户相似度并使用相似用户进行个性化检索的技术作铺垫，提供一定的理论知

识和技术准备。后面将重点介绍这两种以社会化标注为基础，计算用户相似度并使用相似用户进行个性化检索的方法。

3 基于用户相似网络的个性化检索方法

3.1 问题描述

在文献[22]中，Xu 等人提出了使用用户相似度扩展用户标注，形成文档对用户的个性化文档属性及扩展用户属性的方法，并证明了扩展后的属性在个性化检索计算文档得分时具有良好的效果。然而由于网络的自由性，在社会化标注系统中，用户都倾向于给出简单明了的标注，所以标签往往都比较简短，这就导致反映用户情感的标签在不同文档上容易出现重复的现象，例如“funny”、“Interesting”等。所以当使用社会化标注计算用户相似度时需要避免这一点对用户相似度产生负面影响。同时，在使用相似用户计算查询用户的扩展用户属性时，只是简单根据相似度将其他用户的标签加入进来。这虽然很好的解决了标注的稀疏性问题，但却也会将其他用户兴趣中不相关的部分带入，对真实兴趣的构架产生影响。

还需要注意的是，扩展社会化标注需要反复对数据进行扫描。这是因为计算用户相似度、扩展标注及检索排序这三个过程都与文档相关，导致需要反复调用用户、文档、标注及其之间的关系。尤其是计算文档对用户的个性化文档属性时，就需要扫描全部用户和其标注过的文档，之后计算扩展用户属性，又需要再次扫描一遍扩展后用户标注的文档。另外在计算过程中，还会生成一些额外的中间文件需要被存储，又要额外占用存储空间，造成多余开销。在真实网络中，存在着繁多的网页、用户，上述问题的弊端会更加明显。

为了解决上述问题，有必要在仅使用社会化标注的情况下，提出一种能在计算用户相似度时突出用户特性，检索时又能在包含到更多网页的情况下不过度增加算法开销的方法，从而提升个性化质量。同时，用户在同一文档上给出类似标注也应该成为用户相似性的重要参考。

本文以使用社会化标注更好地判断用户相似度以及使用更简单的方法使个性化检索覆盖到用户标注过以外的更多文档为目标，提出了一种基于用户相似网络的个性化检索算法（*SUR-PR*）。算法将用户给出的全部标注类比于传统信息检索中一篇文档，通过使用已经被证实有效的文档表示形式对标签进行处理以突出用户特性，并引入了用户在共同标注过的文档上的相似性这一概念。在进行文档得分计算时，将用户之间的相似关系看作网状关系，并考虑节点之间的相互影响，以相似用户对文档的兴趣扩充查询用户对文档的兴趣，从而减少查询开销、提升检索质量。该方法包括两步：

（1）使用文档表示方法表示用户给出的标注并融入用户在相同网页上的相似度对用户相似度进行计算；

（2）将相似用户对文档的兴趣得分加入查询用户对文档的兴趣得分中后融入文档得分计算公式中，从而使用户对文档的兴趣得到扩展。

下面的章节将详细介绍算法的流程及效果。

3.2 基于社会化标注的用户相似度计算

本文认为，在表示用户属性时，单纯使用用户给出标签的频率当作权重无法准确表达用户在想法上的特性。这是因为文档中会存在一些感情化的标签，虽然出现次数较多但与文档内容关系很小，其会将真正能代表用户特色、出现次数较少的标签掩盖掉。所以本节从用户属性表达上出发，寻找更准确地计算用户相似度的方法。

3.2.1 用户属性表达

在传统信息检索领域，已经有不少文档的表示方式被证明为是简单有效的。TF-IDF 就是其中比较常用的一种。其通过 TF 表示词语在文档中起到的作用，为文档频率；IDF 表示词语在所有文档中所起的作用，为逆文档频率。两者相乘就能很好地表达词语对于区分文档的作用。这是因为只有文档频率高，而逆文档频率低的词语能具有较高的分值，这样的词语不仅对于文档具有代表性，同时也能在区分文档时起到较大的作用。TF-IDF 的文档表示方式因其有效性在信息检索中被广泛地采用，简单的思想也使其可以较容易被扩展到其他资源之上。

假设将用户给出的全部标签看作一个文档集合，则 TF-IDF 模型可以在其上得到很好的应用。标签对于用户的重要程度通常可以使用其在用户全部标注中出现的比例来表示。对于同一用户而言，其给出频率越高的标签越能代表用户的兴趣。所以基于这个比例计算的 TF 值能很好的代表标签与用户的关系，其中 n_{uj} 表示用户 u 给出标签 j 的次数：

$$tf_{u,j} = \frac{n_{uj}}{\sum_k n_{k,j}} \tag{3.1}$$

仅考虑标签对于用户的重要程度无法完全代表相应标签对于区分用户之间兴趣不同的能力，这是因为同样的标签可能被多个用户使用过，从而导致因为部分标签相同而具有较高的用户相似度。因此，本文以 IDF 为基础，加入根据使用标签的用户数决定大小的逆用户频率 IUF：

$$iuf_i = \log \frac{|U|}{|\{j : t_i \in u_j\}|} \tag{3.2}$$

其中$|U|$为网络中的用户总数，$|\{j:t_i \in u_j\}|$为使用过标签t_i的用户数目。可以看到，使用某一标签的用户数越多，则标签对于区分用户兴趣不同的贡献也就越小，计算所得值也越小，反之则越大。

将两者相乘，就得到了 TF-IUF 公式，利用此公式（3.3），用户属性中越属于用户特性的标签将具有越大的权重，而那些共性的标签将具有较小的权重，用户之间将得到更好的区分。

$$tfiuf_{u,j} = tf_{uj} \times iuf_i \tag{3.3}$$

3.2.2 用户相似度计算

在上述处理后，用户属性对于用户兴趣的代表程度得到了增强，因为共性标签的权重经过计算下降，而特性标签权重上升。此时使用改良后的用户属性计算用户相似度，能得到一个比较理想的结果。但仅如此依然不够，这是因为用户在标注网页时比较自由，不同用户浏览并标注过的文档必然存在一定的不同，所以其标签也必然因为浏览的文档不同而有所不同。所以实际上，用户之间的相似度会受到用户浏览过文档的影响，单纯考察其给出的所有标签的相似度只能较好的标识出在标签上类似的用户，而无法做到对看过同样文档但总体属性具有较大差异性的用户进行识别和判断。

基于这个考虑，本文认为，除了考虑用户在整体标签上的相似度以外，还应该考察用户在浏览过的相同文档上标签的相近程度。由于文档之间是相互独立的，所以在考察用户在文档上的相似度时，应该在每一个文档上独立考虑。所以，本文认为，两个用户在文档的相似度应按照公式（3.4）计算：

$$DSim(u,u') = \frac{1}{|D_u \cap D_{u'}|} \sum_{i=1}^{|D_u \cap D_{u'}|} Sim(v_{u,d_i}, v_{u',d_i}) \tag{3.4}$$

其中v_{u,d_i}表示用户 u 对文档 d_i 给出的标注，v_{u',d_i}表示用户 u'对文档 d_i 给出的标注，$D_u \cap D_{u'}$表示两个用户标注过的文档的交集。为了更准确的表达用户之间相似的情况，应该把其与两个用户的用户属性相似度进行融合。比较简单又实用的想法是，两个用户共同标注过的文档越多，其在文档上的相似度对总体相似度的贡献应该越大，这是因为两个用户在越多文档上相似，则其越可能相似。基于这个原因，本文选择将两种相似度进行差值求和，差值的权重取决于两个用户标注过的文档中相同的比例，计算方法如公式（3.5）：

$$PerSim(u,u') = \frac{|D_u \cap D_{u'}|}{|D_u|} \cdot Dsim(u,u') + \left(1 - \frac{|D_u \cap D_{u'}|}{|D_u|}\right) \cdot Sim(p_u, p_{u'}) \quad (3.5)$$

根据上式进行计算时，系数能较好地控制两种相似度在融合时的权重，注意到分母使用了$|D_u|$而不是$|D_u \cup D_{u'}|$，这是因为本文认为计算目标用户相似度时，使用与目标用户标注相同文档的比例更能合理地表达两种相似度对用户相似度的贡献程度。计算之后，只要用户在标注过的相同文档上具有较高相似度，其整体相似度也会得到提升，可以解决用户的相似度被整体掩盖的情况。

在相似度计算完成之后，为了保证后续检索的质量，需要舍弃一些相似度较小的用户。所以需要设置一个阈值 *T*，只有相似度大于 *T* 的用户才会被判断为相似用户，从而提升整体质量。

3.3 融合相似用户推荐的文档排序方法

在个性化检索中进行文档打分和排序时，通常需要考虑两个方面，一方面是查询与文档内容的相关程度，另一个方面是文档与用户兴趣的相关程度。所以本文的得分计算公式以 Bouadjenek 等人的工作为基础，对非社会化部分保持不变，而社会化部分进行改良扩展。

本文将用户之间的相似关系看作网状结构，用户和文档实际上形成了一个庞大的关系网，这样的网可以用一个无向图来表示，图中的每一个节点都为用户，而边则为两个用户之间的相似关系，边的权重可以即是两个用户的相似度。这样形成的一个无向图 <*G*,*E*>（其中 $G \in U$）对于扩展用户对文档的兴趣很好的价值，计算文档和用户的相似度也可以从这个无向图出发，进行计算。

对于每一个用户而言，都对一定的文档存在兴趣，这种兴趣可以用文档被标注的标签与用户兴趣的相似度进行表示，即每一个用户节点都附着一些用户感兴趣的文档。由于用户之间的相似度代表用户兴趣的相近程度，所以越相近的用户其对于文档的兴趣也可能越相似，这就说明可以使用相似用户对文档的兴趣来对目标用户的进行扩充。所以在个性化检索时，文档与用户之间的兴趣实际上是文档与用户和其相似用户共同作用的结果。同时与用户和相似用户兴趣相关的文档能得到更高的分数，这一做法会使排序时对用户兴趣的考量更为准确。同时注意到，相似用户可能标注过目标用户未标注的文档，这一做法也使目标用户未看过的文档与其产生了兴趣关联，使用户有兴趣的文档范围得到了扩展。具体计算公式如（3.6）所示：

$$R(d,q,u)=\alpha\cdot\left[Sim(p_u,p_d)+\sum_{u'\in U_T}PerSim(u,u')\cdot Sim(p_{u'},p_d)\right]+(1-\alpha)\cdot\left[\beta\cdot Sim(q,p_d)+(1-\beta)\cdot Score(q,d)\right] \quad (3.6)$$

从公式中可以看到，两个用户之间如果相似度越高，其对于文档的反馈的权重也越高，而由于通过设置阈值控制了相似用户的相似度，所以可以认为社会化部分的得分相对是合理的。对于目标用户而言，相似用户的意见有助于其更好地找到与自己兴趣相关的文档。

同时应该注意到的是，采用此文档得分计算公式时，由于避免了社会化标注扩展的步骤，使用户对于文档的标注仅需在得分计算时扫描一次，并且不会有额外的中间信息生成，所以在空间和时间上都会有节省。

3.4 实验设计

本节主要评价了本章算法的表现，并将其与最相近的 *D-PR* 进行比较。因为本章的主要目的在于证明两者在个性化得分部分上的不同对整体搜索结果的提升，所以令 β=1 以去除 *Score*(*q*,*d*)进行简化，从而去除基于网页内容的非个性化得分部分的影响。

3.4.1 数据集

本文实验所用数据集中的数据都来自 Del.icio.us 网站，与文献[51]所使用的数据集相同，都为 CABS120k08。数据集中不仅包含有网页 URL 及其上标注过的用户和使用的标签，每个 URL 还有一个网页分类信息。数据集的具体参数见表 3.1 所示：

表 3.1 CABS120k08 全部统计数据

Tab. 3.1 Overall statistics of CABS120k08

统计内容	分类信息	所占比例
总文档数	117434	
总类别数	84663	
总用户数	388963	
总标签数	4673134	唯一标签：35.4%
被分类文档数	117434	100%
被标记文档数	59126	50.3%

从其中可以知道，数据集中的文档都进行了专业的分类，并且大部分网页都存在至少 1 个标签。虽然对比于真实环境、该数据集的数据量并不是很大，但根据其统计数据，可以认为其适用于本章的实验。

为了保证实验质量，本文去除了所有未标注过任何文档的用户以及未被任何标签标注过的文档。筛选后的数据集共包含 388963 个用户，59126 个文档和 3647266 个标签。之后所有的数据再经过两步处理：（1）对所有标签词进行词干化处理，保证有相同意义的标签能够聚集在一起；（2）对标注进行切词：一些标签可能由多个常用词语组成，例如“*java&programming*”、“*java_programming*”等，这些标签虽然计算机无法识别，但往往也包含一定的语义信息，因此将这些词语切分开来有助于后续对于标签的处理。

3.4.2 评价方法

在现今个性化检索的研究中，还未存在一种通用的、被普遍认可的评价方法。这是因为对比于传统信息检索仅需考察文档排序的先后关系，个性化检索的结果评价往往还与提出查询的用户相关，即结果的评价是需要用户参与的。而在实验中，真实用户的参与往往是受条件限制的，所以达不到很好评价结果的水平，故对于个性化检索结果的评价是困难的。

基于上述考虑，一些研究从同样是检索资源的社会化标注出发，寻找可能的评价方式。事实上，用户在网络中的标注行为与其在网络中的检索行为正好可以看作互逆的关系[52-53]，两者具有很强的关联性。由于查询与标注都来源于用户，选择标注文档和选择查询结果也为类似过程，所以如果一个用户在一篇文档上给出了标注，则当该用户用同样的标注作为查询且检索的结果中包含这篇文档时，用户有很大可能会浏览。所以在判定检索结果时，可以将用户用查询标签标注过的文档认定为相关文档。

所以，本文与前人一样，选择使用用户的标注作为查询以检验个性化的效果。在选择查询时采用从数据集所有的三元组(u,t,d)中随机选择的方法。为了与实际情况相符，只选择标签 t 项长度为 2-4 个单词的三元组。每组实验选择 100 个三元组作为查询，对于一组中的每一三元组(u,t,d)，看作是用户 u 使用标签 t 作为查询，目的是搜索到文档 d。实验中，每次共进行 10 组，每组实验在选择完 100 个三元组后将这些三元组剔除，使用剔除后的数据集作为后续个性化检索的基础，以减小实验结果受到的影响。将 10 次实验结果的平均值作为方法的最终检索性能，以保证实验的准确性和有效性。

由于对于每个检索而言，仅有一篇文档会被认定为相关文档，而其他文档的相关性无从判断，所以在评价检索效果时，采用平均排序倒数值（*MRR*），该指标对于一组实验中每一个文档排序的得分结果为 $1/r$，并在一组实验的全部结果上取平均，即：

$$MRR = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{r_i} \tag{3.7}$$

其中 r_i 为在第 i 个查询中，相关文档在个性化检索结果中的位次，n 为一组实验中的查询总数。

3.5 实验结果及分析

3.5.1 阈值选择

为了保证个性化检索的质量，在进行计算时只希望选用相似度较高的用户，为此设置了阈值 T 对用户相似度进行筛选，以确保用户质量。由于阈值 T 的设置需要根据实际情况进行，因此首先同样讨论其大小问题。对阈值取不同的值进行计算，得到结果随着阈值的增加先上升后下降的情况，而阈值在 0.2 附近的结果如图 3.1 所示：

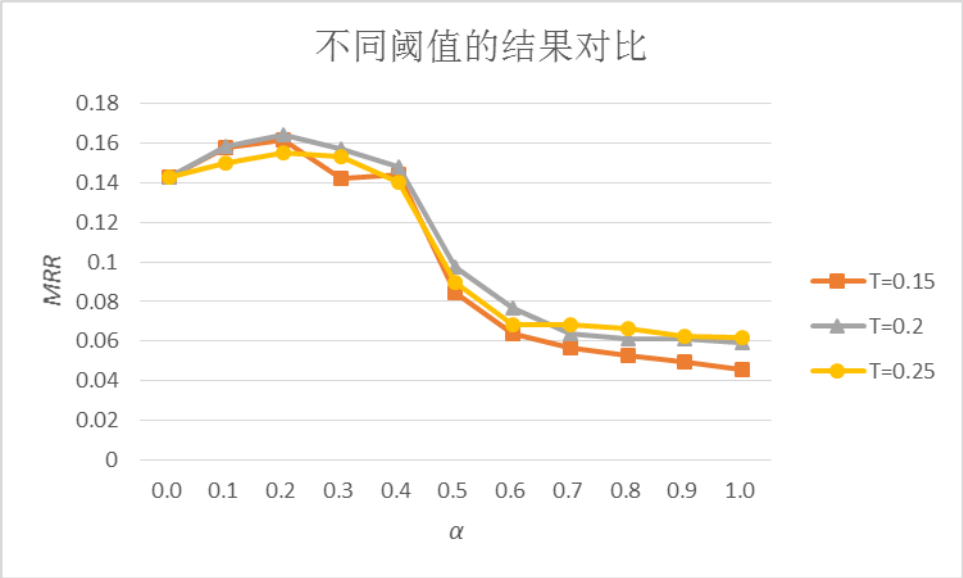

图 3.1 不同阈值下的个性化检索结果

Fig. 3.1 The result of different threshold

从图中可以看到，阈值为 0.2 时，个性化检索取得了最好的效果，所以后续与 Baseline 作比较时取阈值 T=0.2。经过分析，阈值降低的原因主要是在计算用户相似度时降低了用户的共性而提升了个性，所以相似度计算的结果普遍会出现一定程度的下降。同时应该注意到，不论阈值多少，结果都随着 α 的增大先上升后下降，并且降至一个比较低的水平。这说明单纯考虑社会化得分或非社会化得分对于个性化检索都无法获得一个较好的结果，只有将两者结合起来，才能改善个性化检索的结果。而 α 等于 0.2

或 0.3 时，取到最大值，可见即使是进行个性化检索，文档和查询的相似程度依然是一个不可忽视的部分，对于信息检索起到决定性的作用，这一点也可以在 α 等于 1 时的结果得到证实。而随着 *T* 的降低，α 等于 1 时的结果呈现下降的趋势，这是因为阈值越低，低相似度的相似用户人数就越多，则进行文档得分计算时，所出现的干扰和噪音也可能越多。在实际使用中，需要根据实际情况调整阈值，以达到理想的效果。

3.5.2 结果对比及分析

在确定阈值之后，为了检验本章所提出的算法的效果，将其与最近的 *D-PR* 及 *SoPRa* 作比较，比较时筛选相似用户的阈值 *T* 选定为 0.2。实验对比结果如图 3.2 所示：

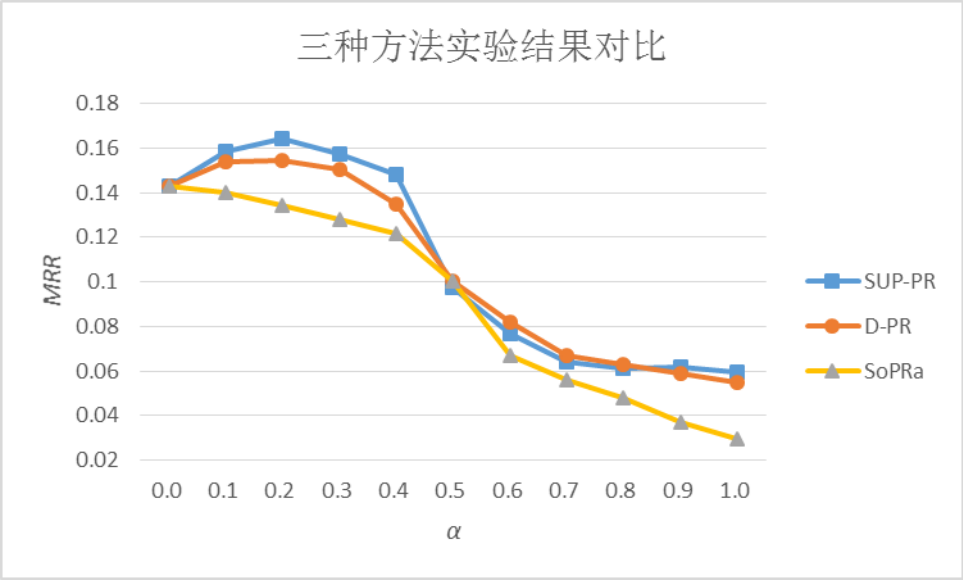

图 3.2 三种方法实验结果的对比

Fig. 3.2 The comparison of three method

从图中可以看到，本章的方法相较于 *D-PR* 表现出了更有优势的结果。其中在 α=0.2 时取到本章实验的最大值 0.1653，比 *D-PR* 在同一个 α 时的最大值 0.1546 提升了 6.92%，这说明本文提出的基于相似用户推荐的个性化检索算法是有一定的优势的，本文经过分析认为这主要是由于采用了更好的方式表示用户属性，在计算相似度时也采用了更为合理的方法，以致相似用户的判断更加准确，从而提升扩展效果。同时在 α 大于 0.5 时，本章方法与 *D-PR* 所得结果差不多，这进一步说明查询与文档的相似程度对于个性化检索而言起着重要的作用。同时，两种对用户兴趣进行扩展的方法对比于不扩展的 *SoPRa* 都有一定的提升，说明用户标注过的网页总是较少的，并不完全能表达用户兴趣，扩展后能将兴趣延伸到更多网页上，对于个性化检索的提升是有益处的。

应该注意到，在完成用户的相似度计算后，本章的方法仅在查询时需要扫描一遍文档，对比于 *D-PR* 在扩展社会化标注及检索时各需扫描一遍文档的做法，在时间上会有缩短。实验结果表明，本章方法对比于 *D-PR* 在相似度计算完成后计算一篇文档的文档得分上，大约能节省 27.4%的时间。考虑到由于数据集较小、文档数较少，所以节省时间的效果并不十分明显，随着文档数量的增多，在时间开销上的优势将更好。同时由于不需要额外空间存储用户的扩展属性及文档对用户的个性化文档属性，所以在存储空间上也有一定的节省。

3.6 本章小结

本章主要探索了在社会化标注系统中基于用户相似网络的个性化检索方法。本章在前人对用户属性的总结上，提出引入信息检索中文档的表示方式，以突出不同用户之间的特性，消减带来不利影响的共性。同时引入了用户在共同标注过的文档上的相似度这一概念，将其融入相似度计算，使得仅使用社会化标注时也能获得较准确的用户相似度。在此基础上，本章从用户之间的相似关系出发，提出将用户看作网络中的节点，相似度为网络中的边，并将网络中节点互相的影响引入文档得分计算公式中，融合目标用户和相似用户的影响共同计算目标用户对文档的兴趣，从而使计算结果能够以较简单的方式覆盖到更多潜在相关的文档，进而提升个性化检索的效果。

本章的实验结果表明所提出的方法对于改善个性化检索的效果具有一定的效果，能够较好地提升信息检索的满意度。同时经分析和验证，该方法在时间和空间上也存在一定的优势。这说明仅使用社会化标注，只要能够用合理的模型表示用户的兴趣、用户之间的关系及用户和文档的关系，个性化检索也能取得不错的效果。

4 融合用户相似度和用户质量的个性化检索方法

4.1 问题描述

基于社会化标注计算用户相似度并扩展标注的方法确实为个性化检索带来了一定的提升。但受限于社会化标注本身具有的问题，单纯使用社会化标注的方法很难达到较高的水平。这是因为：一方面，在互联网中，用户总是自由地对资源进行标注，所以不同用户的标注往往处于不同的层次，部分专家用户的标注准确可靠，而部分用户的标注意义不大，在扩展时不考虑标注的质量会给扩展的用户属性带来噪音。另一方面，用户在网络中浏览的文档总会存在一定的差异性，这时单纯只考虑其在标注上的相似程度并不完全，因为可能出现用户因为兴趣不同而在不同类型网页上给出相同标注的情况，所以单纯的用户属性相似度并不能满足扩展要求。这些都会对个性化检索的结果产生负面影响，需要在计算时予以避免。

以表 2.1 所示社会化标注为例，假定 *Carl* 提出了查询 *q*=“Interesting Film”，首先统计各个用户的用户属性如表 4.1 所示：

表 4.1　用户属性向量

Tab. 4.1　User profile

用户	English	Comedy	Interesting	Boring	Chinese	Action
Alice	2	2	2	1	1	0
Bob	0	0	2	1	0	1
Carl	0	1	2	1	0	0
David	0	1	2	1	2	1

按照公式（2.15）计算 *Carl* 与其余三位用户的相似度：

$$\begin{aligned} Persim(Alice, Carl) &= \frac{7}{\sqrt{14}\cdot\sqrt{6}} = 0.764 \\ Persim(Bob, Carl) &= \frac{5}{\sqrt{6}\cdot\sqrt{6}} = 0.833 \\ Persim(David, Carl) &= \frac{6}{\sqrt{11}\cdot\sqrt{6}} = 0.738 \end{aligned} \tag{4.1}$$

可以看到，与 *Carl* 相似度最高的用户为 *Bob*，假定将筛选相似用户的阈值设定为 *T*=0.8，则只有 *Bob* 被选择为相似用户，这与前面分析得出的事实不符。

事实上，这是因为单纯使用标签相似度并不能准确代表用户之间相似的程度。*Bob* 与 *Carl* 的兴趣其实正好相反，相似的仅仅是给出过的标注。将其用于后续计算时，会因为偏好不同而引入错误的信息，所以对用户标注的扩展也会出现偏差，而最后的个性化排序也必然无法取得良好的效果。

为了解决上述问题，在基于用户相似性的个性化检索方法中，需要提出一种能够更好地判断相似用户，更好地扩展社会化标注的方法，用来提升用户扩展属性的准确度，提高个性化检索的效果。因此，本文在文献[22]的基础上，对基于用户相似度的个性化检索方法进行了更深入的研究。

本文以更准确地判断用户相似程度以及提高扩展标注的质量为目标，从用户标签本身以及用户标注 w 的行为中进行分析，将用户的兴趣和质量与其对于文档的选择及对文档的想法进行关联，认为只有选择和想法都相同的用户才是相似用户，而扩展标注时不同用户的质量会影响扩展的准确度。也就是说，看过类似网页并且个给出类似标注的用户才应该为相似用户，而相似的高质量用户的标签对于扩展有更高的价值。因此在计算用户相似度时，应该同时考虑用户属性相似度及用户标注过的文档类别相似度，在扩展时应该同时考虑用户之间的相似度及用户的质量。于是，本文综合考虑用户标注过的网页的分类及给出的标签，提出了一种融合用户相似度和用户质量的个性化检索方法（*PRuFaC*）。

融合用户相似度和用户质量的个性化检索方法包括以下三个部分：

（1）计算用户相似度：本文在使用用户属性计算相似度的基础上，加入了用户标注过的网页类别的相似度，使相似度的计算不仅考虑用户在想法上的类同，也考虑用户标注网页所属类别的相似性。该方法能够综合筛选出标注过类似网页并且有类似想法的用户，以提升用户相似度判断的准确性。

（2）计算用户质量：本文受 Guo 等人[49]启发，提出了一种基于社会化标注和网页分类的用户质量算法(User Quality Mining, UQM)。该算法能够根据用户标注的标签、文档及文档所属类别等信息为用户计算一个相对合理的质量分数，以此确定用户给出的标签所具有的可信性，方便后续对社会化标注等的扩展。

（2）扩展社会化标注并计算文档得分：将用户相似度与用户质量相结合，对用户的标注进行较高质量的扩展，从而得到较好的文档对用户的个性化文档属性及扩展用户属性，从而更加准确地对文档计算得分并排序。

下面的章节将详细介绍算法的流程及效果。

4.2 网页分类知识介绍

4.2.1 网页分类

为了使用户浏览网页更加方便快捷，网络中逐渐出现了一些对网页进行分类的方法。其通过专人浏览网页之后，将网页分到预定的类别下，使互联网中的网页都拥有比较准确的分类。现今网络存在很多种网页分类的方式，其中 ODP[48]（Open Directory Project）是公开的人工网页分类记录中最大、最广泛的一种。其中的 590000 多个类别涵盖了超过 4000000 个节点，并由超过 65000 名志愿者进行维护。由于这些分类是由专人进行的，所以具有一定的可信性。这样的分类信息本身就对网页进行了一个划分，所以在其上进行检索时，可以使用分类缩小范围，这使信息检索变得更加容易，也能够提升检索的效率。

网页分类信息一定程度上可以算成来源于“专家”，其数量有限但相对可信度较高。这种分类是对网页内容的高度概括，并将其与类似网页放在一起，将无序的网页组织成了一个有序的层状结构，这就使网页之间也产生了一定的联系。例如，表 4.2 给出了表 2.1 所述的 5 个文档的分类信息，可以看到 5 篇文档都属于“Film”类别，都与电影相关，但 d_1、d_3 与戏剧相关，d_1、d_3 与动作电影相关、d_5 与恐怖电影相关。这样的分类说明 5 篇文档当中存在共性，但也存在一定的差异性。这种专业化的分类对于信息检索而言是非常有效的资源。

表 4.2　文档分类示例

Tab. 4.2　Example of document categories

文档	分类信息
d_1	Film/Comedy
d_2	Film/Action
d_3	Film/Comedy
d_4	Film/Action
d_5	Film/Horrible

4.2.2 网页分类的结构和特性

由于本文部分工作使用了 ODP 的网页分类，所以介绍网页分类时，都以 ODP 为例。

ODP 为树状结构，其中有 16 种被广泛认可的顶级类别，每一种顶级类别下都按层次分布着若干层类别，每一个层级都比上一层级的分类更加具体细致。比如某体育网页的分类为“Sport\Ball\Basketball”就是一个具有三个层级的分类信息，其中 Sport 为第一

层级，比较笼统概括；Ball 为第二层级，将第一层级进行了细化；而 Basketball 为第三层级，进行了更加具体的分类。ODP 中的每一个节点都具有如上所示的分类信息，并且大部分节点的分类信息的层级都超过 3 级，以便准确地对网页进行区分。

网页分类信息最大的好处在于其是由相对专业的用户完成的，所以其可信度超过一般的社会化标注，并且对于收录的每一个网页，其都有一个所属的类别。这就使在进行信息检索的研究时，分类信息可以作为一个普遍资源加入进去。而其专业化的分类又使检索不用再在全网络中搜索，可以有目的地进行。用户与这些分类的网页之间的关系也成为了待挖掘的一块资源，用户喜欢的网页分类通常能较好地反映用户的兴趣，是个性化检索的有力根据。

4.3 基于社会化标注和网页分类计算用户相似度

本文认为，社会化标注来源于用户对文档的思考，是用户在浏览过网页后心中的想法，所以事实上其并不能完全代表用户的兴趣。而用户在标注浏览的文档时，总是偏向于标注其感兴趣的文档，所以可以认为用户选择标注的文档本身一定程度上也包含了用户的偏好信息。而网页的分类信息因为其较好的专业度和较高的可信性，能较好地区分不同用户选择文档的类别和习惯，所以网页分类也应该被纳入兴趣和偏好的构建中。将每个用户标注过的网页类别按频率进行统计，就可以得到相应用户感兴趣的网页的类别分布，本文将之称为用户类别属性。

但如果仅单独考虑用户类别向量的相似度显然也是不完全的，这是因为用户只有在浏览过后才会对文档有个准确的判断，所以看过相同文档的用户也可能给出不同甚至是完全相反的标注，此时自然也不能判定两者为相似用户。所以本文认为，在利用 VSM 模型计算两个用户在标注和类别上的相似度之后，应将二者相乘，用来表示两者在兴趣和偏好上的相似度。计算公式如（4.2）所示：

$$Persim(u', u) = Sim(c_{u'}, c_u) \times Sim(p_{u'}, p_u) \qquad (4.2)$$

其中，c_u 代表目标用户的类别属性，由对用户 u 标注过的所有网页的类别进行频率统计得到；$c_{u'}$ 为用来计算相似度的其他用户的类别属性，同样是由对用户 u' 标注过的所有网页的类别进行频率统计得到。这样只有用户同时具有较高的属性相似度和较高的类别相似度时，才会具有较高的用户相似度，也就是标注过越多类似文档并且具有越类似标签的越是相似用户。

再看上面的例子，为了方便说明，统计用户类别向量时使用分类信息的第二层级，用户的类别属性如表 4.3 所示：

表 4.3　用户类别属性向量

Tab. 4.3　User category profile

用户	Comedy	Action	Horrible
Alice	2	2	0
Bob	1	2	0
Carl	2	1	0
David	1	1	1

可以看到，*Carl* 与 *Bob* 所标注过的网页存在一定的差异性，而 *Carl* 与 *Alice* 标注过的网页比较相似。此时结合用户属性及用户类别属性，按照公式（4.2）计算用户相似度：

$$\begin{aligned} Persim(Alice, Carl) &= \frac{6}{\sqrt{8}\cdot\sqrt{5}}\cdot\frac{7}{\sqrt{14}\cdot\sqrt{6}} = 0.725 \\ Persim(Bob, Carl) &= \frac{4}{\sqrt{5}\cdot\sqrt{5}}\cdot\frac{5}{\sqrt{6}\cdot\sqrt{6}} = 0.667 \\ Persim(David, Carl) &= \frac{3}{\sqrt{3}\cdot\sqrt{5}}\times\frac{6}{\sqrt{11}\cdot\sqrt{6}} = 0.572 \end{aligned} \tag{4.3}$$

注意到与 *Carl* 相似度最高的用户从 *Bob* 变为了 *Alice*，而从前面对表 2.1 的分析中我们也可以知道，*Alice* 与 *Carl* 同样喜欢喜剧，并且两者也标注过相似的网页，这说明计算结果与真实情况是相符的。能获得这样的结果是因为其不仅在与兴趣相关的标注文档的选择上与 *Carl* 相似，也在代表想法的标签上反映出了与 *Carl* 的相通之处。这也说明本文的方法确实能更准确地计算用户相似度并判断相似用户，为后续的扩展带来更好的效果。

这里注意到用户的相似度普遍出现了下降，这是由于对比于仅使用用户属性的 *D-PR*，本文在计算相似度时乘上了两个用户的类别属性的夹角余弦值，而用户标注过的网页总是存在一定的差异，所以相似度的值会进一步下降。考虑到这一点，在设置筛选相似用户的阈值 T 时，需要下调一定的水平，以保证相似用户筛选的质量。

4.4　基于社会化标注和网页分类的用户质量计算

4.4.1　用户质量定义

用户质量(User Quality)指用户在网络中的行为所具有的质量，通常可以用来代表用户的专业性和可信性。由于用户质量与其在网络中的行为有密切关系，所以可利用该用户在网络中所留下的记录的可靠性来衡量，也可以说是用户在网络中行为的准确程度。

在社会化标注系统中，通常使用交互增强理论来挖掘用户质量：交互增强理论的规则表明用户、资源和标签三种实体之间存在着紧密联系，高质量标签和高质量资源能生成高质量用户；高质量用户和高质量标签能生成高质量资源；而高质量用户和高质量资源也能生成高质量标签。这就说明在社会化标注系统中三者有着明显的迭代关系，而通过这种迭代关系进行计算，就可以挖掘得到各个用户的用户质量。

4.4.2 用户质量算法

本小节将主要介绍本文所使用的文档用户质量算法。由于交互增强理论在用户、文档和标签三种实体上的迭代关系，使得用该理论计算用户质量是简便易行的。因为标签可以融入文档和用户的关系之中，所以质量分数可以看作仅在用户和文档间进行转移。基于 Guo 等人工作[49]的启发， SocialPageRank 算法中三种实体的转化关系可以简化为用户和文档二者的关系，而标签则用来提供对于用户与文档的转移计算的支持。于是，本文中用户和文档间的质量分数的迭代计算公式如（4.4）所示：

$$\begin{cases} U^{k} = M_{DU}^{T} D^{k-1} \\ D^{k} = M_{UD}^{T} U^{k} \end{cases} \tag{4.4}$$

在上述迭代公式中：U^k 和 D^k 分别表示经过第 k 次迭代后用户的质量分数和文档的质量分数；M_{DU} 和 M_{UD} 分别表示文档到用户和用户到文档的转移矩阵，两种实体的质量分数交替改变，最终将达到一个较稳定的收敛值。SocialPageRank 算法的收敛性在前人工作中已经得到了证明，所以与其基于相同框架提出的简化算法也可以用类似方法证明收敛，故而使用此简化方法计算用户质量是可行的。

与 Guo 等人不同的是，本文认为，用户在标注文档时，其标注过越多同类别的文档，越可能是该领域的专家。因为其对这一领域进行了更多的研究和发掘，所以其在这一领域的认知也更为专业。在初始计算时，这样的用户应该拥有高认知分数，对文档的标注做出了更多的贡献。因此，本文在计算用户质量的过程中，将用户标注文档的类别分布也考虑在内，具体执行办法如下：

（1）对标注过文档的每一个用户提取出其用户类别属性；

（2）将对应用户标注过的与该文档相同类别的文档数量赋予该用户作为认知分数。

如果将用户标注同类文档的数量直接作为对于这类文档的认知分数，那么显然其会随着标注同类文档数量的增加而直线上升，这显然与事实不相符。事实上，这种用户的认知分数应该会随着用户看过同类文档的数量上升而减慢上升的趋势，标注的文档越

多，减慢的趋势也越明显。所以本文使用公式(4.5)对用户的认知分数进行控制，从而反映出上述趋势，而非无休止的上升。

$$Rs(u,d)=\sqrt{\left|\left\{d_u,d_u\in C(d)\right\}\right|} \tag{4.5}$$

其中，d_u 表示用户 u 标注过的文档。公式中平方根的处理能有效抑制标注文档数量偏多的用户带来过大的影响，控制在相对合理的程度。

与此同时，假设有 N_u 个用户和 N_d 个资源，令 M_{UD} 表示从用户到文档的转移矩阵，为 $N_u \times N_d$ 维；M_{DU} 表示从文档到用户的转移矩阵，为 $N_d \times N_u$ 维。转移矩阵中的每个元素 $M_{UD}(u_i,d_j)$ 和 $M_{DU}(u_i,d_j)$ 由如下公式计算得到:

$$M_{UD}(u_i,d_j)=\frac{\sum_{t_k\in T(u_i,d_j)}\frac{1}{2^k}\cdot p(t_k|d_j)}{\sum_{u\in U(d_j)}\sum_{t_k\in T(u,d_j)}\frac{1}{2^k}\cdot p(t_k|d_j)} \tag{4.6}$$

$$M_{DU}(u_i,d_j)=\frac{\sum_{t_k\in T(u_i,d_j)}\frac{1}{2^k}\cdot p(t_k|u_i)}{\sum_{d\in D(u_i)}\sum_{t_k\in T(u_i,d)}\frac{1}{2^k}\cdot p(t_k|u_i)} \tag{4.7}$$

其中，$p(t_k|d_j)$ 表示文档 d_j 生成标签 t_k 的概率，这一概率可以看做标签对文档描述的可信程度；$p(t_k|u_i)$ 表示用户 u_i 生成标签 t_k 的概率，这一概率表示了用户给出标签的可信程度。这两种生成概率都可以根据用户标注的行为按下面的公式计算得到：

$$p(t|d)=\frac{\sum_{u\in U(t,d)}Rs(u,d)}{\sum_{t_x\in T(d)}\sum_{u\in U(t_x,d)}Rs(u,d)} \tag{4.8}$$

$$p(t|u)=\frac{\sum_{d\in D(u)}p(t|d)}{\sum_{t_x\in T(u)}\sum_{d_j\in D(u)}p(t_x|d_j)} \tag{4.9}$$

注意到在计算文档生成标签的概率时，各个用户给出的标注在求和时都乘以了用户在这一类别上的认知分数，从而将用户标注过的文档类别情况及相应标签融合进了迭代的过程中。

迭代之后，每个用户都将拥有一个相对合理的质量分数来代表其标注行为所具有的质量和可信程度，并用 $QS(u)$ 代表用户 u 的质量分数。获得所有用户的质量分数后，在

使用相似用户计算文档对用户的个性化文档属性及扩展用户属性时，就可以将用户质量加入进去从而提升扩展质量，最终达到提升个性化检索效果的目的。

4.5 社会化标注扩展及文档得分计算

4.5.1 社会化标注扩展

在前人的工作中，对社会化标注进行扩展以解决标注稀疏性，改善个性化检索效果的方法已经得到了证明。这说明只要利用合理的形式使用户标签能够覆盖到更多的文档，个性化检索的效果能得到很好的提升。

在扩展社会化标注时，前人工作中认为与查询用户兴趣越相似的用户对于扩展带来的价值越大，所以其通过阈值筛选出相似用户后，使用基于相似度的权重进行扩展。这与普遍的认知是一致的。

本文认为，在用户兴趣之外，用户质量也会对扩展的效果产生影响。在实际网络中，用户总是偏向于相信专家的语言，认为其所说的话具有相对较高的可信度。在社会化标注系统中，标注就可以看作用户的“语言”，这些“语言”的可信程度就取决于给出标注的用户的质量。所以本文在扩展时，不仅考虑用户相似度，也将用户质量纳入扩展公式中，提出了基于相似用户质量的扩展方式，如公式（4.10）所示：

$$p_{u,d} = \sum_{i=1}^{|U_d \cap U_T|} \left(V_{u_i,d} \cdot Persim(u_i, u) \cdot QS(u_i) \right) \quad (4.10)$$

考虑到用户更容易接受与自己近似的用户的标签，而不是单纯质量高的用户的标签，所以在选择扩展用户时，依然设置一个阈值 T，只有相似度大于 T 的用户才会被用来进行扩展，而小于 T 的用户即使有较高用户质量也直接舍弃。同时由于用户之间的相似度以及用户的质量都与标签的可信度直接相关，都能影响标签被用户接受的可能性，所以公式中将二者相乘，以保证二者的共同作用。

4.5.2 文档得分计算

传统的信息检索通常使用查询和文档的相似度作为文档得分，这种方法不考虑提出查询的用户，在检索时总是返回相同的结果。这样的结果虽然在过去的时代取得了良好的效果，但随着用户对于信息检索的要求不断提高以及对个体的重视，搜索引擎也需要根据不同用户的情况不同而返回更加准确的检索结果，这既是个性化搜索。

在个性化搜索中，通常为了根据用户的喜好返回不一样的检索结果，选择对用户的兴趣和偏好进行挖掘后融入检索模型之中，从而使文档得分能受到用户兴趣的影响。在

这一方面，常用的有两种模式：一种是将用户的兴趣因子直接加入文档打分公式中，使文档得分直接受影响从而改变排序结果；另一种是在搜索引擎返回初始结果后，基于用户兴趣对文档进行重排序，以达到检索结果根据用户兴趣进行反馈的目的。

在这一方面，本文与前人一样，认为文档得分应该由两部分组成：1、文档与用户兴趣的相似程度；2、查询与文档的相似程度。前一部分代表与用户兴趣相关的社会化得分，后者为与文本内容相关的非社会化得分。由于考虑到在对结果进行个性化重排序时，查询与文档的相似程度依然时文档检索中一个重要的因素，因此两部分得分应该综合进行文档得分的评定。

基于以上原因，本文在计算基于社会化标注的文档得分时，采用了前人的计算公式：

$$R(d,q,u)=\alpha\cdot Sim(p'_u,p_{u,d})+(1-\alpha)\cdot[\beta\cdot Sim(q,p_d)+(1-\beta)\cdot Score(q,d)] \quad (4.11)$$

从公式中可以看到，社会化和非社会化两部分得分共同作用之下生成了文档的最终得分，α 的大小能够控制两部分的分数对于最终文档得分的贡献，可以通过调节其值的大小改变得分最终得分中基于用户兴趣部分的重要程度。

4.6 实验设计

本章由于同样是研究社会化部分的得分改进情况，所以也令 β=1 以简化公式。

4.6.1 数据集

本章所使用数据集与第三章相同，为来自 Del.icio.us 网站的 CABS120k08。同样为了保证实验质量，去除了所有未标注过任何文档的用户以及未被任何标签标注过的文档。处理后的数据情况如表 4.4 所示：

表 4.4 数据集概况

Tab. 4.4 Overall statistics of dataset

用户数	网页数	标注词
388963	59126	3647266

之后将所有的数据同样经过两步处理：（1）对所有标签词进行词干化处理，保证有相同意义的标签能够聚集在一起；（2）对标注进行切词：一些标签可能由多个常用词语组成，例如“*java&programming*”、“*java_programming*”等，这些标签虽然计算机无法识别，但往往也包含一定的语义信息，因此将这些词语切分开来有助于后续对于标签的处理。故最后本章使用的数据，实际上与上一章完全相同。

4.6.2 评价方法

由于本章的工作最后实际上也反映在对文档检索结果的排序上，故要检验方法的效果也只能从排序结果上下手。基于同样的考虑，本章使用跟上一章相同的评价方法。即：如果一个用户在一篇文档上给出了标注，则当该用户用同样的标注作为查询检索的结果中包含这篇文档时，用户有很大可能会浏览。所以同样使用上一章的评价方法对本章实验结果进行评价。

同样从三元组(u,t,d)中采用随机选择的方法，选择标签 t 项长度为 2-4 个单词的三元组。每组实验选择 100 个三元组作为查询。实验中，每次共进行 10 组，每组实验在选择完 100 个三元组后将这些三元组剔除。最后将 10 次实验结果的平均值作为方法的最终检索性能，以保证实验的准确性和有效性，评价指标依然采用平均排序倒数值（*MRR*）。

4.7 实验结果及分析

注意到对于数据集中每一个文档，其网页分类都存在若干个层级，而不同层级所具有的数量和详细程度也各不相同。而这一不同会对用户之间的相似度及用户质量产生影响。为了验证分类信息的不同层级对检索结果的影响，本文分别选择了第一层级的网页分类（$PRuFaC^1$）和第二层级的网页分类（$PRuFaC^2$）分别进行实验。

4.7.1 阈值选择

对于阈值的选择，本文将两种分类层级的方法分开进行考察，对其在不同阈值 T 下进行实验，得到的结果如图 4.1 所示：

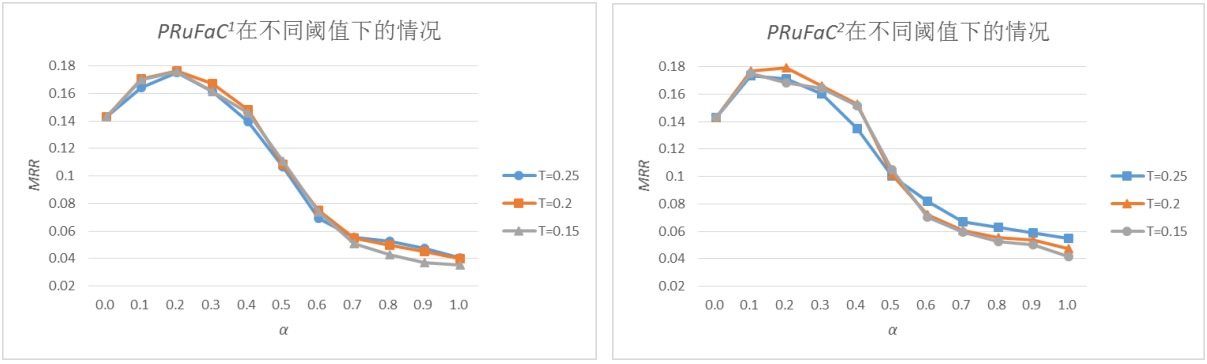

图 4.1 *PRuFaC* 在不同分类层级下的阈值情况

Fig. 4.1 Different threshold of *PRuFaC* in different level of category

从图中可以看出，不管使用哪一层级的类别，两种方法都在 T=0.2 时取得最好的效果，并且 *PRuFaC*^{2} 的最大值稍好于 *PRuFaC*^{1}，本文认为这是因为使用第二层级的分类信息时，对用户的兴趣和质量的判断将会更加准确。同时由于加入前两级中的任一级分类对检索结果的影响相差不大。因此，将 T=0.2 的 *PRuFaC*^{2} 与一些较相近的方法进行比较，其中包括 *UP-PR*[11]和 *SoPRa*[12]。比较时为了验证本文加入用户质量的效果，也将其与仅计算用户相似度而不加入用户质量的结果进行比较。而不管使用哪一个层级的分类信息，其随着 α 的变化趋势是相似的，同时在效果上，两个层级也相差不多。所以在于 Baseline 进行比较时，仅使用 *PRuFaC*^{1} 作为代表进行比较。

4.7.2 结果对比及分析

为了验证本章所提出的方法的有效性，将其与 *D-PR* 和 *SoPRa* 的效果进行对比，结果如图 4.2 所示：

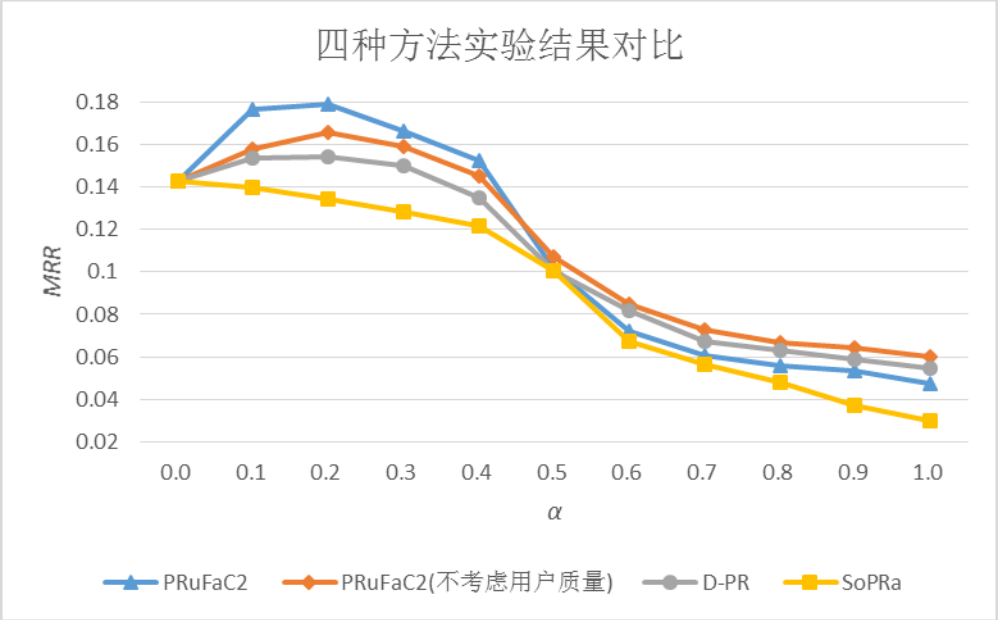

图 4.2 两种方法实验结果的对比

Fig. 4.2 The comparison of two method

可以看到，本文的方法相较于 Baseline 中最好的 *D-PR* 在最大值上表现出更好的结果。其中 *PRuFaC*^{2} 在 α=0.2 时取到实验的最大值 0.1789，比 *D-PR* 在同一个 α 时的最大值 0.1546 提升了 15.7%。同时，几种方法的对比也可以给出以下信息：（1）在不加入用户质量时，本文的方法对比于 *D-PR* 仍然有一定量的提升，这说明了本文考察用户相似度的方法比单纯使用社会化标注进行计算得到的结果更加准确，而加入用户质量之后

的提升则说明用户质量对于检索质量有较大影响；（2）不加入用户质量时，α从 0 到 1 的结果都优于 Baseline，而加入用户质量后其在 0.5 开始比 *D-PR* 下降更快，最后的结果也略差，这进一步说明用户质量对于扩展社会化标注的影响是较大的；（3）对比于 *SoPRa*，三种方法都有较大优势，这说明扩展社会化标注的方法对于改善个性化检索的质量有一定的效果，而除 *SoPRa* 以外的三种方法在 α从 0 到 0.2 的过程中都不断的上升，说明扩展的方法能对文档排序带来正面的影响；（4）当 α大于 0.2 时，三种方法都呈现下降的趋势，这表示过度依赖个性化的部分会得到一个很不理想的结果，基于网页内容的相似度在个性化检索里依然起到了非常重要的作用。这个结果在一定程度上说明本文基于社会化标注和网页分类信息计算用户相似度和用户质量后扩展社会化标注的方法对于改善个性化检索的结果有一定的优势。

4.8 本章小结

本章探索了综合使用网页分类信息和社会化标注进行个性化检索的可能性。首先提出了一种结合两种资源计算用户相似度的方法，以确保准确判断出相似用户；之后使用两种资源对于用户质量进行判断并计算出用户质量的得分；最后在此基础上进行社会化标注的扩展及文档得分计算和排序。该方法不仅考虑了对用户潜在兴趣的挖掘，同时也考虑了不同质量的用户带来的影响。

本章的实验证明了统筹用户两个方面进行个性化检索的效果，证明该方法对于总体效果有较大的提升。这说明单独使用社会化标注进行个性化检索，对于用户兴趣的构建层次比较低，而将其与其他资源联合，可能达到更加理想的效果。

结　论

个性化检索技术在当今这个时代已经得到了长足的发展，其使用的资源从内部到外部，非常丰富。随着 Web2.0 的普及，使用社会化标注进行个性化检索研究已经得到了很大的关注。本文基于这一背景下展开研究，具体来说，本文工作可以概括为以下两个方面：

首先，本文首先考察了在仅使用社会化标注的情况下如何改善个性化检索的效果，。本文将用户给出的标签与文档进行类比，提出了使用 VSM 模型的方法表示用户属性，以消减共性，突出个性。并且引入用户在共同标注过的文档上的标签相似性，使相似用户的判断更为准确。之后为了减小社会化标注的稀疏性造成的影响，同时降低查询时所耗时间，将用户之间的相似关系看成网状结构，提出了节点间相互影响的文档兴趣得分计算方法，使用户对于文档的兴趣能在相对简单的方式下较准确地覆盖到更多文档。本文实验验证了这种方法对于个性化检索的效果有一定的提升，并同时说明了其在资源消耗上的优势，这表明对用户标签的合理利用对于个性化检索效果的提升是有效的。

其次，为了更好地改进检索质量，本文还提出了一种融合用户相似度和用户质量的个性化检索方法。该方法在使用社会化标注的基础上，加入了分类信息作为额外资源。首先使用两种资源分别代表用户兴趣与偏好的一个侧面，结合起来计算不同用户之间的相似度，以使相似用户的判断更为准确。之后基于两种资源对用户的质量进行计算，得到用户质量分数，并基于相似用户及其质量对社会化标注进行扩展。最后使用扩展得到的文档对用户的个性化文档属性及扩展用户属性计算文档得分并排序。本文实验表明，使用分类信息的第一层级或第二层级，在个性化检索的结果上差异不大，都能较好地改善个性化检索的效果。这说明，在扩展社会化标注时，准确考虑用户兴趣及用户的质量对于扩展的效果能带来比较大的影响。由于本文使用的方法相对简单，这一方法还存在较大改进空间。

值得注意的是，本文提出的个性化检索算法所涉及的思想都相对简单，引入分类信息时使用的方法也比较直观。在个性化检索中，还存在着更多的资源和更多的方法等待挖掘，这也将是日后本文的研究方向。本文提出的两种基于社会化标注的个性化检索方法对于结果均表现出了一定的提升，这表明社会化标注是个性化检索的良好资源，对于其更深层次的挖掘和利用将为个性化检索提供更加有力的支持和帮助。

参 考 文 献

[1] Lambiotte R, Ausloos M. Collaborative tagging as a tripartite network[C]. Proceedings of the International Conference on Computational Science, Reading, UK, 2006:1114-1117.

[2] Laura G M. Social Bookmarking, Folksonomies, and Web 2.0 Tools [J]. Searcher, 2006, 14(6):26-38.

[3] Kiu C, Tsui E. TaxoFolk: A hybrid taxonomy-folksonomy structure for knowledge classification and navigation[J]. Expert Systems with Applications, 2011, 38(5):6049-6058.

[4] Lu D, Li Q. Personalized search on Flickr based on searcher's preference prediction[C]. Proceedings of the 20th international conference companion on World wide web, New York, NY, USA, 2011:81-82.

[5] Andrews P, Zaihrayeu L, Pane J. A classification of semantic annotation systems[J].Semantic web,2012,3(3):223-248.

[6] Borrego A, Fry J. Measuring researchers' use of scholarly information through social bookmarking data: A case study of BibSonomy[J]. Journal of Information Science, 2012, 38(3):297-308.

[7] Xie H, Li Q, Mao X. Context-Aware personalized search based on user and resource profiles in folksonomies[J]. Web Technologies and Applications, 2012, 7235:97-108.

[8] Doerfel S, Zoller D, Singer P, et al. How social is social tagging[C].Proceedings of the companion publication of the 23rd international conference on World wide web companion, Republic and Canton of Geneva, Switzerland, 2014:251-252.

[9] Gupta M, Li R, Yin Z, et al. An overview of social tagging and applications[J]. Social Network Data Analytics,2011:447-497.

[10] Bouadjenek M R, Hacid H, Bouzeghoub M, et al. Using social annotations to enhance document representation for personalized search[C]. Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, NY, USA, 2013:1049-1052.

[11] Berger F C, van Bommel P. Personalized search support for networked document retrieval using link inference[C]//Database and Expert Systems Applications. Springer Berlin Heidelberg, 1996: 802-811.

[12] Speretta M. Personalizing search based on user search histories[D]. University of Kansas, 2000.

[13] Liu F, Yu C, Meng W. Personalized web search for improving retrieval effectiveness[J]. Knowledge and Data Engineering, IEEE transactions on, 2004, 16(1): 28-40.

[14] Pretschner A, Gauch S. Ontology based personalized search[C]//Tools with Artificial Intelligence, 1999. Proceedings. 11th IEEE International Conference on. IEEE, 1999: 391-398.

[15] Gauch S, Chaffee J, Pretschner A. Ontology-based personalized search and browsing[J]. Web Intelligence and Agent Systems, 2003, 1(3): 219-234.

[16] Chirita P A, Nejdl W, Paiu R, et al. Using ODP metadata to personalize search[C]//Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2005: 178-185.

[17] Ma Z, Pant G, Sheng O R L. Interest-based personalized search[J]. ACM Transactions on Information Systems (TOIS), 2007, 25(1): 5.

[18] Joachims T, Granka L, Pan B, et al. Accurately interpreting clickthrough data as implicit feedback[C]//Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2005: 154-161.

[19] Chirita P A, Firan C S, Nejdl W. Personalized query expansion for the web[C]//Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007: 7-14.

[20] Shen X, Tan B, Zhai C X. Implicit user modeling for personalized search[C]//Proceedings of the 14th ACM international conference on Information and knowledge management. ACM, 2005: 824-831.

[21] Xu S, Bao S, Fei B, et al. Exploring folksonomy for personalized search[C]//Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2008: 155-162.

[22] Xu Z, Lukasiewicz T, Tifrea-Marciuska O. Improving Personalized Search on the Social Web Based on Similarities between Users[M]//Scalable Uncertainty Management. Springer International Publishing, 2014: 306-319.

[23] Shen X, Zhai C X. Exploiting query history for document ranking in interactive information retrieval[C]//Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM, 2003: 377-378.

[24] Sugiyama K, Hatano K, Yoshikawa M. Adaptive web search based on user profile constructed without any effort from users[C]//Proceedings of the 13th international conference on World Wide Web. ACM, 2004: 675-684.

[25] Luxenburger J, Elbassuoni S, Weikum G. Task-aware search personalization[C]//Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2008: 721-722.

[26] Teevan J, Dumais S T, Liebling D J. To personalize or not to personalize: modeling queries with variation in user intent[C]//Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2008: 163-170.

[27] Haveliwala T H. Topic-sensitive pagerank[C]//Proceedings of the 11th international conference on World Wide Web. ACM, 2002: 517-526.

[28] Wu X, Zhang L, Yu Y. Exploring social annotations for the semantic web[C]. Proceedings of the 15th international conference on World Wide Web, Edinburgh, Scotland, 2006:417–426.

[29] Uddin M N, Duong T H, Nguyen N T, et al. Semantic similarity measures for enhancing information retrieval in folksonomies[J]. Expert Systems with Applications, 2013, 40(5):1645-1653. [30]

[30] Ye Z, Huang X J, Jin S, et al. Exploring social annotation tags to enhance information retrieval performance[M]//Active Media Technology. Springer Berlin Heidelberg, 2010: 255-266.

[31] Bouadjenek M R, Hacid H, Bouzeghoub M. LAICOS: an open source platform for personalized social web search[C]. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, Chicago, Illinois, USA, 2013:1446-1449.

[32] Rawashdeh M, Kim H, Alja'am J M, et al. Folksonomy link prediction based on a tripartite graph for tag recommendation[J]. Journal of intelligent information System, 2013, 40(2):307-325.
[33] Lllig J, Hotho A, Jaschke R, et al. A comparison of content-based tag recommendations in folksonomy systems[J]. Lecture Notes in Computer Science, 2011, 6581:136-149.
[34] Noll M G, Meinel C. Exploring social annotations for web document classification[C]//Proceedings of the 2008 ACM symposium on Applied computing. ACM, 2008: 2315-2320.
[35] Plangprasopchok A, Lerman K. Exploiting social annotation for automatic resource discovery[C]//AAAI workshop on Information Integration from the Web. 2007: 86-91.
[36] Hotho A, Jaschke R, Schmitz C, et al. Information retrieval in folksonomies: search and ranking[C]. Proceedings of the 3rd European Semantic Web Conference, Budva, Montenegro, 2006:411–426.
[37] Bao S, Xue G, Wu X, et al. Optimizing web search using social annotations[C]. Proceedings of the 16th International World Wide Web Conference, Banff, Alberta, CANADA, 2007:501–510.
[38] Bouadjenek M R, Hacid H, Bouzeghoub M. Sopra: A new social personalized ranking function for improving web search[C]//Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2013: 861-864.
[39] Noll M G, Meinel C. Web search personalization via social bookmarking and tagging[M]. Springer Berlin Heidelberg, 2007.
[40] Robertson S E, Walker S. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval[C]//Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. Springer-Verlag New York, Inc., 1994: 232-241.
[41] Vallet D, Cantador I, Jose J M. Personalizing web search with folksonomy-based user and document profiles[M]//Advances in Information Retrieval. Springer Berlin Heidelberg, 2010: 420-431.
[42] Jones K S. A statistical interpretation of term specificity and its application in retrieval[J]. Journal of documentation, 1972, 28(1): 11-21.
[43] Micarelli A, Gasparetti F, Sciarrone F, et al. Personalized search on the world wide web[M]//The adaptive web. Springer Berlin Heidelberg, 2007: 195-230.
[44] Xu Y, Wang K, Zhang B, et al. Privacy-enhancing personalized web search[C]//Proceedings of the 16th international conference on World Wide Web. ACM, 2007: 591-600.
[45] Culliss G A. Personalized search methods: U.S. Patent 6,539,377[P]. 2003-3-25.
[46] Carmel D, Zwerdling N, Guy I, et al. Personalized social search based on the user's social network[C]//Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009: 1227-1236.
[47] Al-Khalifa H S, Davis H C. Exploring the value of folksonomies for creating semantic metadata[J]. International Journal on Semantic Web and Information Systems, 2007,3(1):13-39.
[48] Open directory project. http://dmoz.org/.

[49] Guo Q, Liu W, Lin Y, et al. Query Expansion Based on User Quality in Folksonomy[M]//Information Retrieval Technology. Springer Berlin Heidelberg, 2012: 396-405.
[50] Veres C. The language of folksonomies: What tags reveal about user classification[J]. Lecture notes in computer science, 2006, 3999:58-69.
[51] Noll M G, Meinel C. The metadata triumvirate: Social annotations, anchor texts and search queries[C]//Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on. IEEE, 2008, 1: 640-647.
[52] Benz D, Hotho A, J äschke R, et al. Query logs as folksonomies[J]. Datenbank-Spektrum, 2010, 10(1): 15-24.
[53] Bischoff K, Firan C S, Nejdl W, et al. Can all tags be used for search?[C]//Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008: 193-202.

攻读硕士学位期间发表学术论文情况

致　　谢

光阴荏苒，从我第一次来到美丽的滨城大连，来到美丽的大连理工大学已经经过了七年时光。在这七年的求学生涯中，我从一个懵懂无知的少年成长成为了一个努力上进、求知若渴的青年。现在到了即将与这段生涯分别的时刻，在这离别失分，我怀着万分的不舍，感激所有在这段路程中给予我关怀和帮助的人。

首先，我要感谢在我成长的过程中给予我最多的父母。从小到大，是他们一直支持着我，为我的学业之路保驾护航。在这一路的奋斗中，他们不光给予我经济上的支持，还在我觉得困扰、遭受挫折时给我最多的温暖。在他们无私的奉献下，我才能顺利完成学业。

其次，我要感谢的是我的导师林鸿飞教授。在研究生的三年里，林老师是对我影响最深远的人。他为人严谨，学识渊博，对本专业有着丰富的经验和了解。在我对于学业存在困惑时，主动传授相关的专业知识，并指导我如何将研究做得更好并易于被他人理解。他为人热情，对待学生也如朋友一般，不仅关心学生学习的情况，更是亲自过问生活的冷暖。林老师不仅是我们研究上的良师益友，更是我们为人处世的良好模范。在林老师的带领下，信息检索实验室拥有了既严谨又活泼的氛围，学生们互帮互助、关系融洽，是求学阶段难得的环境。

再次，我要感谢实验室的其他老师，平时的大小事务都是许侃老师在细心管理、悉心照顾，我们才能拥有如此良好的科研环境。林原老师平常经常在实验室给予我们直接指导，在我们遇到苦难时也经常是他第一个为我们答疑解惑，指明下一步的方向。还有孙晓琳老师，对我的研究工作提出了不少的修改意见，使我在对科研的认知上更进一步。这些老师都具有很深的科研功力，是实验室的楷模。这些老师就如实验室的明灯，在我前进的道路上时刻为我照亮前方。

最后，我要感谢我实验室的马云龙师兄和徐博师兄，作为实验室的博士师兄，他们是对我提供最多直接帮助的人，对我的研究总是尽心尽力，为我解决了很多平日里的困难。另外，还要感谢已经毕业的常天舒师姐、张平师兄、刘文飞师兄、姚兰师姐、郭青师姐等，他们努力奋斗的身影都是我学习的榜样。

同时，我还要感谢在求学路上有一群志同道合的好伙伴，吴雨、郝辉辉、吴慧、赵明珍、杨阳、刘敏捷、孙东普、王亮以及其他实验室的学弟学妹们。我们是科研上的好同伴，生活中的好朋友，感谢你们陪伴我度过有意义的三年，让我的生活充满色彩。

我要感谢这一路遇到的每一个人，因为你们才成就了今天的我。

感激你们，谢谢！

大连理工大学学位论文版权使用授权书

本人完全了解学校有关学位论文知识产权的规定，在校攻读学位期间论文工作的知识产权属于大连理工大学，允许论文被查阅和借阅。学校有权保留论文并向国家有关部门或机构送交论文的复印件和电子版，可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印、或扫描等复制手段保存和汇编本学位论文。

学位论文题目：基于社会化标注的个性化检索方法研究

作 者 签 名 ：_____ 日期：____年____月____日

导 师 签 名 ：_____ 日期：____年____月____日