SHEHUIHUA BIAOQIAN DE
YUYI JIANSUO YANJIU

社会化标签的语义检索研究

宣云干◎著

出版社

社会化标签的语义检索研究

宣云千　著

东南大学出版社
·南京·

内 容 提 要

社会化标注系统已发展成为重要的网络资源组织与共享平台，是 Web 2.0 以来网络服务的重要发展方向之一。本书鉴于绝大多数用户的标注行为符合社会共同认识，存在基本的、潜在的语义结构支配标签的出现和资源语义构成，结合潜在语义分析这一信息检索代数模型，提出基于潜在语义分析的标签语义检索模型和排序算法，来解决由于标签的模糊性、不规范及资源数据庞大等造成的漏检、低效率等问题。

本书在大量真实数据集上进行了实验研究，验证了方案的实用性与可行性，可作为系统开发者、信息研究者、网络服务提供者、信息管理者、高等院校师生及相关人员学习、研究的参考书。

图书在版编目(CIP)数据

社会化标签的语义检索研究 / 宣云干著. —南京：东南大学出版社，2013.9

ISBN 978-7-5641-4453-1

Ⅰ.①社… Ⅱ.①宣… Ⅲ.①网络检索—研究 Ⅳ.①G354.4

中国版本图书馆 CIP 数据核字(2013)第 205539 号

社会化标签的语义检索研究

出版发行	东南大学出版社
出 版 人	江建中
社　　址	南京市四牌楼 2 号(邮编：210096)
经　　销	全国各地新华书店
印　　刷	北京京华虎彩印刷有限公司
开　　本	700 mm×1000 mm　1/16
印　　张	10.5
字　　数	172 千字
版　　次	2013 年 9 月第 1 版
印　　次	2013 年 9 月第 1 次印刷
书　　号	ISBN　978-7-5641-4453-1
定　　价	45.00 元

本社图书若有印装质量问题，请直接与营销部联系，电话：025-83791830。

前　言

作为社会性的个人，通过社会化软件，可以建立社会网络和自组织群体，构建社会关系，而这样的社会关系所具有的重要价值，正被越来越多的人认识到。社会化软件突出了个体自主性地参与和发挥，帮助人们在浩瀚的互联网中发现富有价值的信息，并在信息整理和共享过程中发现生活的乐趣和更多的同好知己。社会化软件还可以成为私人的搜索引擎，所得出的结果有时候比 Google 搜索还要准，而且是经过过滤的东西，其含金量自然更高。

多用户通过因特网共享自己的资源固然是好事，但是这种信息组织方式过于自由了，资源的分类也十分随意，因此造成了大量的重复信息和含义模糊的信息，对用户检索和利用有价值信息极为不便，查全率、查准率都得不到保障。例如，由于用户的认知程度不同，对同一事物的揭示存在很大的差异，造成“一词多义”和“多词一义”的现象，导致检索时彼此完全没有关系的内容会聚集在一起，或相关联的内容并不全出现在检索结果中。而且，在社会化软件中，一般资源都很庞大，单凭标签的定位与比较会造成检索结果过多的情况，就是提升检索的速度也没有办法改善整体的检索效果。使用者很希望能直接找到自己需要的东西，以减少过滤大量检索结果的负担。

因此，近年来研究者们非常关注提高社会化软件的检索效率问题。本书以社会化标注系统为例，针对以上的检索弊端，对传统检索算法进行改进，提出基于潜在语义分析的标签语义检索模型，将潜在语义分析技术应用于社会化标注系统的语义分析中，构建基于向量的多维空间语义模型。改进传统 T－R 矩阵的权重计算方法，提出资源全局权重和标签全局权重的概念及计算公式，将局部权重（某个标签在某个资源中的权重）、标签全局权重（标签在所有资源中的权重）

以及资源全局权重(资源对标签权重所能提供的信息量)相结合对权重计算进行改进,更贴切地反映了标签标注频数在标注系统权重计算中的作用,得到新的资源模型。

本书还结合寻找标注用户相似性和资源标注的时序特性,对检索排序算法进行改进,提出排序算法模型,让越贴近的结果越在前面显示。最后通过大量数据集合的实验,验证新方案的实用性和有效性。

本书在编写过程中得到了南京大学工程管理学院朱庆华教授的指导和帮助,同时研究过程中参考了很多前人的研究成果,如果由于本人疏忽遗漏了一些参考文献,还请原作者谅解。希望本书能帮助普通网络用户高效地管理与利用信息,帮助信息管理部门更好地监测热门话题,帮助商业机构更好地跟踪用户群体兴趣转移,同时也希望成为相关专业研究人员学习、研究的参考书。

作　者

2013 年 5 月

目　录

1 绪 论

1.1 研究背景及意义

标签(Tags)是关键词、范畴名称或元数据,在本质上,一个标签仅是一组自由选择的文本关键词[1]。标注(Tagging)是定义标签的过程,是使用者为一项资源加入描述性元数据的工作,是人们在阅读和使用资源时出于需要而留下各种类型的批注的行为[2]。近年来,随着网络的普及,人们在不断创造信息的同时也按照自己的需求组织各类信息,于是出现了大众分类(Folksonomy)的概念。它是网络资源的利用者为了组织和利用某一数字资源而自主地对该资源赋予标签的过程或结果,是一种自下而上的分类方法[3]。古老而传统的标注行为因此而逐步演化成一种社会性的信息资源组织和利用方式——社会化标注(Social Tagging),标注具有了新的应用领域和使用价值。

社会化标注在标注者和被标注对象(资源)之间直接建立了关联,同时通过资源关联到资源作者,通过使用行为关联到资源的使用者(指非标注者部分)[4],由此形成了一种复杂的关系网络——社会化标注系统(Social Tagging System),它存储标签到一个公开网站,并且以关键词来标注它们。要利用社会化标注系统进行标注,使用者必须在一个社会化标注网站进行注册,以存储标签,加入使用者选择的标签,并且表明个别标签为公开或私有[5]。在这个关系网络中,结点具有多进多出的特点,并通过标注之间的聚合(共享语义)在标注者内部、资源内部、资源使用者内部形成同类群,通过任何一个入口都可以对整个网络进行检索和跟踪。

社会化标注系统是目前 Web 的主流发展方向之一,虽然研究时间不长,但取得了阶段性成果。社会化标注表达了用户对资源的感知度与认同度。从标注的数量可以说明被标注对象的受关注度,从标注的时间跨度可以说明被标注对象受关注的持久度,而其他用户对标注的响应度(阅读数和回应数)、标注者在所探讨领域的权威度等也都可以作为评价被标注资源的依据。

社会化标注系统的产生就是为了使用户能够更好地组织和检索信息，虽然开始只是为了方便个人对信息的管理和检索，但由于大量用户的交互，表现出了足够的社会效应，具有了相当的稳定性。标签和其他元数据一样，都是信息检索系统的操作对象，因而它可以作为信息检索的一个新的途径，尤其是检索网络热点时。

社会化标注系统采用的是根据用户自定义的标签进行描述、分类和检索，这就造成了标签的模糊性。由于用户认知程度不同，对标签词义的理解各异，导致对同一事物的揭示存在很大的差异。如在同一标签下，由于同字不同义，可能会出现彼此完全没有关系的内容都出现在检索结果中；由于同义不同字，使得相关联的结果并不全出现在检索结果中。标签的模糊性和不规范性，可能会产生大量“噪音”，或者在检索过程中没有按照检索目标的关联性缩小范围，或者检索结果缺乏有效的关联性排序，关联度高的结果无法拥有较优先的顺序而加重系统负担。这些都会降低检索的准确性，造成检索的困难。

而且，在社会化标注系统中，一般网络资源都很庞大，单凭标签的定位与比较会造成检索结果过多。在过多的检索结果产生之后（一个标签的查找结果会达到几万条，用时只有零点几秒），使用者则必须自己逐个地对搜索结果进行浏览，从中挑选出自己想要寻找的资源。但想要的东西可能并没有包含在反馈的结果中，在这样的情况下提升检索的速度并没有改善整体的检索效果，因此使用者很希望直接找到自己需要的东西，即提高检索的准确性，以减少过滤大量检索结果的负担。因为人们没有精力和时间来遍历所有的检索结果，一般只会查看前几十个排列。

在社会化标注系统中信息检索是根据用户的查询请求，在大规模标注资源集中找到与查询相关的资源。对于社会化标注系统来说，多用户通过因特网共享自己的资源固然是好事，但是这种信息组织方式过于自由了，资源的分类也十分随意，这就产生了大量的重复信息和含义模糊的信息，对用户检索和利用有价值信息极为不便，查全率和查准率都得不到保障。因此，近年来研究者们非常关注提高社会化标注的检索效率问题，最热门的就是社会化标注系统的语义检索研究。

语义（Semantic）是一个语言学上的概念，它是语言文字能够对应于实体概念的体现。简单地说，有意义的语言文字信息就形成了语义。人类对事物的习

惯是以语义的方式来思考，人们希望在社会化标注系统中对资源赋予语义。这可以透过大众分类法，在社会化标注系统中让使用者所做的标签与检索结果之间拥有更多的关联性与概念上的相似度，让越来越多的人开始喜欢并使用社会化标注系统。

语义检索是一种超越传统的、通过定义文档和查询中的概念来进行检索的检索方法[6]。最早的语义检索研究者是 Raphael[7]，他建立了 SIR 系统，该系统分解不同的查询/问题进不同的子程序进行处理。同样 Li 等也着手利用语义信息来获得问题分类器[8]。国内也有学者对语义检索概念进行了定义。余传明认为语义检索是对检索条件、信息组织以及检索结果显示赋予了一定语义成分的一种检索方式[9]；张雷认为语义检索是一种在获得了被检索的数据或信息的语义的基础上，通过对语义进行明确的表示和处理来使得结果在意义上而不仅仅是在语法或结构上满足搜索需求的系统或方法[10]。

从以上定义我们可以看出，语义检索本身还是信息检索，但它更加强调“语义”，这是与传统关键词检索相区别的。语义检索使用语义标注方法半结构化地组织信息，把传统意义上的检索转变为对知识库的推理，挖掘隐含知识。例如，用户要查询的是“操作系统”，则“Windows”也是与之相匹配的词语。基于知识和语义匹配的语义检索在提高检索的查准率和查全率方面都有很好的表现。语义检索作为下一个检索研究的热点，与传统的检索技术相比，它能提高检索的精度和覆盖率，减少不相关的返回结果。

在社会化标注系统中进行标签语义检索，首先要为资源赋予语义，即语义标注。由于目前人工智能的发展还没有达到机器能够完全理解资源内容的程度，所以这个过程是目前语义检索中最能体现出“语义”的环节，也是最困难的过程。在社会化标注系统中，存在某些基本的、潜在的语义结构支配标签的出现和资源的构成，认为一个包含语义的资源出现在以标签为维度的空间中，其分布不是绝对随机的，而是服从某种语义结构；同样的，一个标签出现在某个资源中也同其他出现在该资源中的标签有密切的关系，而非随机出现的。资源是由语义组成，而标签又要放到资源中去理解，这体现了一种“标签-资源”双重概率关系。如果能够将这种语义结构提取出来，归纳标签间的语义联系，就能为信息检索提供一种语义匹配的新方法。

所以如果能找到一种方法来自动获取标签间的语义关系，将标签和资源以

可计算性高和可操作性强且代表语义的形式表示和存储，对标签语义检索意义重大。潜在语义分析(Latent Semantic Analysis，简称 LSA)正是这样一种具有以上特点、可用于语义检索的自然语言统计模型。LSA 使用了强有力的充分自动的统计方法，揭示出标签间和资源间的联系，创立了一个语义或概念空间，利用该空间可实现标签和资源的语义匹配索引以及提取信息。

贝尔实验室的研究显示，LSA 提取信息一致率超过关键词方法 20%～30%[11]。基于 LSA 的语义检索方法以一个标签与资源相联系的大规模矩阵开始，自动地建造了一个语义空间，使得使用者能够发现相关信息。即使提问式中没有任何词与之相关，只要在概念上与该资源的主体思想联系相一致，在语义空间中它们仍然紧靠在该资源附近。因此标签和资源在语义空间的位置可以用来作为一种语义指引。提取信息的过程就是利用提问式中的关键词来识别空间的一个点，在这个点附近的资源按标签向量与资源向量之间点乘的余弦值的大小排列，即按关键词与资源相关程度排列返回给使用者。

从理论意义上讲，基于潜在语义分析的标签语义检索技术的研究促进了当前互联网技术的发展。Web 2.0 及以前的互联网没有可以提供机器可读的语义信息，这种缺陷限制了计算机自动分析处理及进一步智能处理的能力。语义网和社会化标注被认为是下一代互联网的研究热点，使计算机和网页之间能够从语义层面上互相理解和沟通。标签语义检索是语义网技术在社会化标注系统中的应用，对此研究可以直接推动社会化标注系统的发展。

从应用意义上讲，基于潜在语义分析的标签语义检索能够弥补传统标签检索的不足。传统标签检索技术采用关键词匹配的形式，不考虑查询请求的具体含义，由于标签的“一词多义”和“多词一义”，越来越不能适应用户的检索要求；而标签语义检索将标签和用户查询转换成计算机能理解的语义概念，从而检索出与此概念相关的、用户真正需要的信息，克服了传统检索的局限性。

1.2 研究现状

1.2.1 社会化标注系统研究

社会化标注是 Web 2.0 环境下信息组织与检索的一种新方法，一般以2004 年社会化标注这一概念首次被 Thomas Vander Wal 提出作为社会化标注系统研究的起点。近几年来，对社会化标注系统的研究主要分为三类[12]。

(1) 理论研究

主要包括标注行为的认知心理学分析[13,14,15]；标注系统自身结构、语义及运作机制[16,17,18,19]；构建精确描述标注系统的形式语言[20]；标注系统与其他知识组织工具的比较分析[21,22,23]。社会化标注系统是互联网的新应用，引起了计算机科学、心理学、统计学等多个学科学者的关注，研究者从不同的学科角度提出了自己的看法和观点，在不同学科领域的交流和碰撞中不断完善相关理论。

(2) 实证研究

主要包括标签、用户、资源关系实证研究[24,25]；标签实证研究[26,27,28]；用户实证研究[29,30,31]。研究者开展研究的入口通常是社会化标注系统的主要构件，包括标签、用户和信息资源及其之间的关系，使用的方法通常是基于数学理论和统计规律，具体实证研究呈现出数据挖掘的性质。

(3) 应用研究

主要包括社会化标注系统的标签排序及推荐研究[32,33,34,35,36,37]；基于社会化标注系统的知识组织工具研究[38,39,40,41]；社会化标注系统的改进与开发[42,43,44,45]；社会化标注系统中的信息分类显示研究[46,47,48,49,50]。这些研究主要致力于丰富用户体验和提高信息组织与检索的效率，研究者希望通过实验与应用研究更有效地帮助用户管理自己的信息。

从以上的文献分析可发现，社会化标注由 2004 年底被提出以来受到了广泛的关注，并在快速演化和升级。随着大众持续的需求和新的技术手段的加盟，它推陈出新的速度要快于元数据，成为目前海量网络资源组织的热点研究方向。

1.2.2 社会化标注系统中标签的语义标注研究

由于社会化标注所存在的标签同义、多义、缺乏层次等不足，影响了社会化

标注效用的充分发挥，并导致目前社会化标注网站中较低的内容重复利用与兴趣共享程度。实际上，引起同义、多义等问题的主要原因是缺乏语义信息。目前研究的主要思路，一方面是借助常识工具增加标签语义；另一方面是从标注系统中提取出涌现语义，提高对标签的正确理解，减少社会化标注系统中的混乱。

(1) 扩展标签语义

目前已有较多的语义工具如 WordNet、Wikipedia 等可以减轻或消除标签存在的一些弊端。WordNet 可以返回标签所属的类，可以利用该信息检查该标签是否与内容属于同一类。

Christiaens 的研究不仅应用本体将标签进行了层级化表示，而且还具体到了上下层级之间的特定关系[51]。Laniado 将相关标签建立了语义层级，进而帮助用户寻找相关的资源。但该方法对解决标签同义较为有效，而对歧义问题则帮助不大[52]。Specia 在对标签进行预处理后运用统计方法分析了标签的共现，并建立共现矩阵来划分标签簇，使用在线的词典以及本体资源将标签绘制成概念、属性以及例子，并确定已绘制标签间的关系[53]。

Marchetti 通过开发一个新的基于语义性的社会化标注系统 SemKey，对当前的标注系统加以概念上的扩展。在该系统中，标签被分为三类关系：hasAsTopcic，hasAsKind，myOpinionIs，用户需要指出其所标注与内容关系属性。同时 Semkey 也通过 WordNet 来减少歧义，Semkey 的思想是为内容附加更多的信息，而不是只有标签[43]。

Nauman 通过已有的软件(The Open Mind Common Sense Project，简称 OMCS)以及 Conceptent，将用户的查询关键词(标签)扩展为几个相同的概念，然后再进行相应的查找，并对查找结果进行打分，进而得到相应的结果[54]。

Ronzano 则将研究提升到了新的高度，将目标标签的邻居标签吸纳进来，建立了一个名为“Syntag”的库，并且认为资源是由概念所组成，而概念又是由关键词(标签)所组成[55]。基于此，通过 Wikipedia 将其中的文章作为资源，文章标题作为概念，再将文章标题与内容中提取的词作为关键词建立了 Syntag 集。该集由概念与关键词所组成，可以消除歧义，更加全面地表达概念。

Haklackin 在语义层面上讨论数种社会化标签系统标注的方法，提出一个标签协作标注和设计的民俗分类法的概念模型。同时比较已有的标签本体，提

出了一个评价标准[56]。

Jose 提出一个基于语义网技术和社会化标注相结合的平台将不同领域细节本体联合。用户可以添加元数据到资源中，同时用户可以协同标注资源。该系统利用联合不同元数据搜索引擎来定位期望的资源，还通过本体和标签提供浏览能力[57]。

借助常识工具增加标签语义的方法主要是借助 WordNet、Wikipedia 等工具来确定标签之间以及查询与标签之间的类属关系。该方法基本上是基于本体的思想利用本体工具，在标注的时候提示需要什么标签，而不是任人自由标注。其不足之处在于：首先是建立任意两词之间的语义关系是困难的，所以尽管 WordNet 有众多专家的参与，但大多数有实际语义联系的词语的关系没有在 WordNet 中体现出来；其次是仍然有许多词语没有被收录到 WordNet 中；再者是实验证明 WordNet 在提高检索查全率的同时降低了查准率，这给实际应用带来很多不方便；最后是这些工具是英文的，不适合中文环境。

(2) 涌现标签语义

Wu 运用概率论方法挖掘潜藏在用户、资源和标签共现频率中的潜在语义[58]，通过将用户的标注行为用一个概率生成模型(Probabilistic Generative Model)加以表示和处理，最终自动地得到标签的涌现语义，实现了同义与多义标签的识别和区分。尽管该方法从社会化标注中提取了涌现语义，但其得到的结构仍旧是平的，没有层级。

Heymann 等通过将标签作为点、标签相似度为边建立相应的无权图，试图将大量的标签转化为可导航的层次结构的分类[38]。将标签按其所标注的资源的次数表示成向量的形式，利用余弦相似性计算不同标签的相似性并给定相应阈值得到标签的相似图，进而得到潜在层级的分类。Begelman 统计了基于资源的标签共现，并利用分离点去除弱关联的标签，将强关联的标签表示成无向权图，运用聚类分析得到层次性[59]。Halpin 对高出现频率标签形成的共现网络进行分析，指出可以利用这些高频标签与其他标签的关系确定目标标签的意义[18]。

Zhou 应用确定性退火(Deterministic Annealing)算法提出了从社会化标签中自动提取出层次性语义的相关模型来有效地反映语义概念和层级间的关

系[60]。Pasquale通过大众分类方法提出了一个支持用户标注资源的新途径[61]，包括：利用概率方法来加速精确决定两个标签的相似度和概括度(Generalization Degree)；提出两个等级结构和两个相关算法，在一个等级里来安排语义相关的标签组，这样可以让用户根据期望的语义粒度显现他们感兴趣的标签，帮助他们发现最能表达他们信息需求的标签。Jun通过函数计算用户曾用标签被选可能性，并对每个备选标签可能性进行评估，推荐排序最高的给用户[37]。

从标注系统中提取出涌现语义的语义标注方法，大都采用共现矩阵、无向权图等传统语义向量分析方法。向量空间模型理论(VSM)被应用到社会化标注系统中，将资源(URL)内容和查询内容表示成标签项及其权重的向量，每个资源 U_i 被表示为一个 n 维的向量，即

$$\boldsymbol{U}_i=(u_{i1},u_{i2},u_{i3},\cdots,u_{in})$$

其中，u_{ij} 代表第 j 个标签在资源 U_i 中的权重。如果把每个标签看做是向量的一维，那么由这些标签集合就构建了一个空间。这个空间在数学上可以表示为标签(T)-资源(U)矩阵(T-U矩阵)。

向量空间模型将自然语言中的每个资源视为以标签为维度的空间中的一个点，每个标签则被视为以资源为维度的空间中的一个点。向量空间模型的缺点是显而易见的：两个不包含共同标签的资源其相关度为0，而没有考虑到词形不同的标签间仍然存在语义关系，VSM用互不相关的向量代表标签这一点本身与人对词语的认知不符，同时形成的向量维度较大，对标签共现次数影响较大。

1.2.3 社会化标注系统中检索的排序算法研究

标签是资源的简要描述，寻找相似的资源可以被认为是对多个标签的搜索。标签间的重合程度越高，网页就越相似。

相关度就是用来判断获取的资源集合对用户需求满足的程度。度量查询与资源相关性的方法有很多种，最常用的有向量点乘、向量余弦、皮尔逊相关系数，另外欧式距离、Manhattan距离、Minkowski距离等向量测度都可作为语义向量间的相关度。社会化标注不仅可以应用于信息资源的相关度排序，还可以改进搜索和排序的质量。目前排序算法研究主要有以下几种。

Hotho 研究了社会化标注的信息检索问题，基于 PageRank 的思想提出了 FolkRank 算法用以计算用户、标签和资源的重要性和特定主题的排序，并比较了 FolkRank 和 PageRank 之间的不同；在进一步的工作中还分析了基于 FolkRank 的一个特定主题的演化趋势，构建了大众分类的模型和 FolkRank 检索机制[62]。即分两步走，第一步将用户、标签和资源的层次图表转化为无向权重三部图，第二步应用不同的排序方法来处理社会化标注的网络和无向图的偏斜结构。

Szekeyl 构建了 UserRank、PageRank 和 TagRank 的模型，并在 delicious.com 中进行试验；同时也在 Gourmetvillage. org 网站和美味书签中进行了 URLCount 和 URLRank、TagCount 以及 TagRank 和 UserRank 检索结果的比较和分析[63]。AlKhalifa 通过测量大众分类和 Yahoo 关键词设置的重叠率以及索引者主观评价两种系统产生的关键词的质量，来对基于相同网站的 Yahoo API 文本语词抽取技术和大众分类进行了评价[64]。

Wu 通过改进的 Hits 算法，将用户、标签、资源作为点、其关联作为边构建网络以探寻专家用户和权威性的资源[65]。Malizia 基于机器常识(Machine Common Sense)方法，提出了相应的改进算法[66]。值得提醒的一点是，关联性是社会化标注系统最为本质的特征，因此在用户选定一个特定标签的情况下，可以通过分析标签的共现与频率，进而提取出一组标签以具体化用户的搜索行为来探索更为有效的搜索算法。

Barrows 将标签作为一种现有搜索手段的补充进行探讨，如将标签、分类和浏览进行集成，三者作为相互补充的手段[67]。Han 将标签与 Google 的配合使用进行了探讨，先由 Google 检索出结果(URL)，然后将这些 URL 在 delicious.com 上找出相应的标签，将最频繁的标签作为关键词利用 Google 组合式地进行检索，以此来验证标签的检索效率[68]。

Vanderlei 研究发现结合关键词和标签的检索比单独的检索结果要理想，但是这些方法在处理上还比较粗略，且资源在很大程度上受到是否有标注的影响[69]。

Yanbe 提出了基于标注次数的 SBrank 算法，认为两个算法可以结合使用以发挥各自的优势[70]。Koutrika 构造了一个框架用于对标注系统和用户标注行为进行建模，并提出一种基于标注可信任声誉(Reputation)的排序方法，实现