用于Web2.0搜索意图理解的共识语义分析关键技术

赵玉丽　张　引　张　斌　高克宁　朱志良◎著

東北大學出版社
Northeastern University Press

用于 Web 2.0 搜索意图理解的共识语义分析关键技术

赵玉丽　张　引　张　斌　高克宁　朱志良　著

东北大学出版社

·沈　阳·

图书在版编目（CIP）数据

用于 Web 2.0 搜索意图理解的共识语义分析关键技术 / 赵玉丽等著. — 沈阳：东北大学出版社，2018.9
ISBN 978-7-5517-1989-6

Ⅰ. ①用… Ⅱ. ①赵… Ⅲ. ①数据检索 Ⅳ. ①G254.926

中国版本图书馆 CIP 数据核字(2018)第 192503 号

出 版 者：东北大学出版社
地址：沈阳市和平区文化路三号巷 11 号
邮编：110819
电话：024-83683655(总编室) 83687331(营销部)
传真：024-83687332(总编室) 83680180(营销部)
网址：http://www.neupress.com
E-mail: neuph@neupress.com
印 刷 者：沈阳航空发动机研究所印刷厂
发 行 者：东北大学出版社
幅面尺寸：170mm×240mm
印 张：6.25
字 数：90 千字
出版时间：2018 年 9 月第 1 版
印刷时间：2018 年 9 月第 1 次印刷
责任编辑：张德喜
责任校对：东 山
封面设计：潘正一

ISBN 978-7-5517-1989-6 定 价：38.00 元

目 录

第1章 引 言

1.1 研究背景

Web 搜索引擎以其简单便捷的使用方法、庞大的信息索引数量与较高的结果质量在我国获得了 79.4%的使用率[1]，成为了网民首选的信息查找工具。然而受到关键字有限的描述能力[2]及用户的使用习惯[3]的影响，查询条件与用户的搜索意图之间存在的“意图间隙(Intention Gap)”[4]，其典型的表现如关键字的同词异义、异词同义及语义粒度不同等现象，使当前的搜索方法在探索性搜索方面很难有效地满足用户需求。针对这种情况，研究人员提出了面向搜索意图的搜索方法，并逐渐地成为了搜索领域研究的热点问题[5]。

Web 2.0 开放的信息发布方式允许用户自由发布与共享信息。这种自由性激发了用户的参与热情，丰富了互联网上的信息，也导致了 Web 信息组织方式的改变。

Web 1.0 时代信息的发布与组织由网站的管理人员完成，并按照各网站自有的分类体系分类。这些分类体系虽然各不相同，但都形成了用户理解该网站信息的基础，为用户直观理解网站内容提供了支撑。而在由普通用户主导信息发布的 Web 2.0 时代，为了方便用户对信息的分类并免于记忆和使用复杂的分类体系，Web 2.0 使用了开放自由的信息分类方法如标签等。该分类方法依赖用户个人对分类概念的理解，不限制使用的词汇，有利于用户自由发布信息。但是，这种分类方法由于基于用户个人对分类的理解，缺少一个构

成信息发布者和信息查找者之间公共语义基础的“基础架构”，使得意图间隙问题变得更加突出。因此，如何针对 Web 2.0 信息开放自由的组织特点，研究有效的搜索意图理解方法成为了有效支持 Web 2.0 环境下的信息检索所必须要解决的问题。

1.1.1 Web 2.0 与标签系统

Web 2.0[6-7]是相对于传统的、相对封闭的 Web 而定义的以用户为中心的、鼓励所有人参与其中的 Web 应用环境。虽然从诞生之日起，Web 就以其参与方式的开放性、信息传播的快速性及获取信息的便捷性而获得了广泛的使用，但由于网络的管理、服务器的维护及网站的创建仍旧存在着较高的技术门槛，使得在传统的 Web 环境下，信息的发布权仍旧掌握在相对少数的网站管理者手中，更多的普通用户则只能被动地接受信息。而 Web 2.0 的诞生及随之而来的以用户为中心的应用设计理念，则通过为用户提供简单、便捷同时又多样化、个性化的信息发布方法，突破了传统 Web 所存在的信息发布上的技术障碍，使普通用户可以任意地发布与共享任何类型的信息。这种自由性极大地激发了用户的参与热情。毫无疑问的，这种广泛的参与性已经令普通用户成为了互联网信息发布的主体之一，并在令互联网信息获得极大丰富的同时，使其能够覆盖过去无法引起少数的信息发布者注意的更加广泛的主题与领域。

然而，Web 2.0 开放的信息发布方式在为互联网带来更加丰富的信息的同时，其不受控制的信息发布方法也为如何有效地组织、索引进而查找这些信息带来了困难。在传统的 Web 环境中，信息的发布与组织主要由网站的管理人员完成，并按照各个网站自有的分类体系进行分类。这些分类体系虽然各不相同，但一般都按照用户能够直观理解的方式进行组织，同时在较长的时间里保持固定。对于一个分类，用户通常可以通过分类名称来判断该类别所代表的分类语义。而即便对于不能直接判断语义的分类，用户也可以通过检视该分类下的信息来理解该分类所代表的信息。这种公开的、易于

理解的同时相对固定的分类体系成为了信息的发布者和使用者之间共同的语义基础，为信息的发布、组织、索引与查找提供了一个一致平台。

而在 Web 2.0 环境下，开放的信息发布方式令普通用户逐渐地成为了互联网信息发布的主体。这种变化在改变互联网信息内容结构的同时，也改变了互联网信息组织的方式：一方面，开放的信息发布方式在令信息所涉及的主题及领域快速膨胀的同时也导致了分类的进一步细化，这便要求使用覆盖面更广、描述能力更强、更加复杂的信息分类方法；另一方面，普通用户通常没有耐心去记忆并使用一个复杂的信息分类方法，这又要求信息的分类方法必须尽可能简单与直观。这种矛盾的需求使得在传统的 Web 环境下所使用的基于固定分类体系的分类方法不再适应 Web 2.0 环境下信息发布与组织的要求。

针对这一情况，为了适应 Web 2.0 信息发布方式所具有的开放、自由的特点，人们研究并使用同样开放且自由的标签系统[8]来实现对 Web 2.0 信息的组织、索引与查找。在标签系统中，发布、共享或再发布信息的用户为信息附加纯文本标签作为元数据信息，并使用这些标签来组织、索引并查找信息。这种信息组织方法不限制用户所使用的词汇，避免了传统的信息组织方法受到的自身所采用分类法的描述能力的限制，使其可以用于分类组织任意主题及类型的信息[9]。并且，相对于传统的将信息唯一的分类到分类法的某个类目下的过程，标签系统采用一种更加直观的多元分类策略，使分类的形成更加的灵活，并可以充分地利用人类对事物认知的直觉力量[10]。最后，由于不会受到给定分类法的限制和干扰，用户可以更好地集中在对信息的分类过程中，并因此可以触发更多方面的标注角度，令分类结果能够更好地反映信息被发布、共享或再发布时的上下文[11]。

这些方面的事实令标签系统具备了大量传统分类方法所不具备的优势，并使其在获得用户大量使用的同时，也获得了研究人员的大量关注。Robu 等[12]指出，用户在使用标签系统时会对标签的语义

形成共识，并且这种共识信息可以被用于发现概念间的关联。Tsui 等[13]则进一步地利用标签系统获得了概念的层次关系。Wu 等[14]的研究结果表明，从标签系统中提取的全局语义模型可以帮助对标签的语义进行消歧并帮助识别具有相似意义的标签，进而帮助搜索并发现语义关联的资源。Cattuto 等[15]证明了标签结果可以被用于进行搜索引擎的查询扩展服务。Gawinecki 等[16]则采用标签结果来辅助进行服务发现。这些多样化的研究证明了标签系统的潜在价值。

在标签系统中，用户将标签作为元数据标记给资源。这一过程包括了 3 种类型的对象：用户、被标记的资源、被用于标记资源的标签，以及关联 3 种对象的一种关系：标记关系。这一结构使得标签系统可以很自然地被描述为一种三部图结构：用户集合 $U = \{u_1, u_2, \cdots\}$ 、资源集合 $R = \{r_1, r_2, \cdots\}$ 、标签集合 $T = \{t_1, t_2, \cdots\}$ ，以及这些节点之间用以表示标注关系的超边集 $\boldsymbol{E}$ [17]。

这种简单的模型只考虑了最低限度的信息，因此只提供了有限的描述能力。为了适应不同的、更加复杂的应用场景，一些研究也提出了这一模型的改进。Gruber[18]认为，当需要同时处理来自不同标签系统的标注数据时，来源不同的数据需要被分别地处理，并提出了一个四元组标签系统定义 Tagging（Object，Tag，Tagger，Source）。Wu 等[14]则将标签行为抽象为一个包含标签、用户、资源以及标注发生时间的四元组。Schmitz 等[19]则在研究应用关联规则方法到标签系统中时，进一步将标签的上下位关系引入了建模中。这些研究结果表明，随着视角的变化，标签系统可以进行不同方式与角度的建模，并可以传达出不同类型的信息。这种多样性也证明了标签系统不仅仅是基于关键字的元数据信息，更可以体现出大量用户对信息的一致观点。

1.1.2 标签系统的基本特征

标签作为标签系统最为重要的核心，其基本特征已经获得了广泛的研究。最为直观的研究标签基本特征的方法是观察标签的使用

情况。Mathes[20]的研究指出，标签的使用频率服从幂律分布，即大部分的标签只被少数的用户在少数的资源上使用有限的次数，相反的却有少数标签被大量的用户在很多场景上广泛地使用。进一步的研究可以发现，那些使用频率较高的标签通常对应着关键的概念或实体，因此在研究和应用中需要被重点地关注。Sen 等[21]在研究标签系统中用户所使用的词汇集的演进的过程中也发现一些标签只在用户个人的书签中出现，是非常特别的，甚至是用户自己创造的词。这些词汇不会被其他用户所使用，它们的使用范围也仅限于用户用于浏览自己的资源。Halpin 等[11]通过深入研究标签幂律分布的形成过程发现，在基本的标注形成后，其他用户会继续使用已经存在的热门标签对资源进行标注，从而形成越来越稳固的幂律分布特征，而这些重复的标注则可视为用户对高频标签的一种认可。

标签的另一个重要的基本特征是其与传统分类方法的区别。Jacob[10]指出，传统的分类(Classification)是将对象严格地划分到某一个类别中，类别之间是没有重叠的。而类似于标签的分类(Categorization)则更灵活地将对象分成组，组内的对象在特定的背景下具备共同特征，同一个对象也可以存在于多个组中。Körner 从标记动机的角度将使用标签的用户分为分类者与描述者，并认为分类者所做的标记是为了方便自身对资源的访问，而描述者则是为了方便他人对资源的访问[22]。在类似的研究中，Nov 等[23]更具体地提出了对应于分类者的组织动机与对应于描述者的交流动机。分类者通常会使用一些个性化的标签，而描述者则会用规范的标签以及很多同义词标签来描述资源。典型的分类者是网络收藏系统的用户，典型的描述者则为博客、视频的发布者。很明显的，规范的标签更易于分析标签间的关系，也因此更有利于标签结果的应用。

标签的根本出发点在于有效地组织、索引并查找资源，因此标签的搜索性能也是其重要的基本特征。Heymann 等[24]针对标签是否能够帮助改善网络资源搜索质量的问题进行了大量的研究。他们的研究发现，对于网络书签资源，标签出现在被其所标记的超过 50% 的资源中，并且只有 20%的标签没有出现在被标记资源、其父链接

资源和其子链接资源中。然而，标签虽然能够提供无法从其他数据源获得的、可以用于搜索资源的信息，但是仅仅依靠标签数量及使用分布等信息仍旧难以形成明显的影响，因此对标签进行更深入的研究是获得良好搜索效果的基础。Stampouli 等[25]认为，标签歧义性是影响资源搜索准确率的重要因素，并设计了借助 Wikipedia 消除标签歧义的方法。Wu 等[14]也研究了标签的歧义识别问题，并且发现从标签系统中提取的全局语义模型可以帮助对标签的语义进行消歧，并识别具有相似意义的标签。

作为标签系统三部图模型的一部分，标签之间的关系也形成了复杂的网络特征。吴等[26]通过复杂网络的分析方法对标签系统中标签间的关系进行了分析，发现基于共现的标签网络具有较小的平均路径长度以及较大的聚类系数，体现出明显的小世界特征。对这种现象的一个合理的解释是网络中有类似于树形结构根节点的标签将众多标签联系了起来，这意味着为标签构建概念层次关系是可行的。贾等[27]对网络书签应用 Del. icio. us 中中文标签的特点进行了全面的分析，发现用户倾向于选择简单的词汇来描述资源，且概括性的词汇的使用多于具体性词汇的使用。由于概括性的标签更适合作为层次关系中的节点，这一现实有利于形成更清晰更有价值的标签层次结构。Zlatic[28]等利用一组拓扑质量指标研究了照片分享网站 Flickr 与文献组织网站 CiteULike 的标签系统所形成的网络，发现这些标签网络具有类似的性质。这一结果使得上述研究可以推广到很多类似的系统中。

标签语义特征研究可以通过对标签进行分类实现。Eda 等[29]认为，标签可以分为主观标签和客观标签，同时只有客观标签可以用于构建标签层次关系。Lin 等[30]则将标签分为标准标签、复合标签、术语标签以及无意义标签 4 个类别。Xu 等[31]指出标签分为如下 5 种类型：基于内容的标签，如 Autos，Honda Odyssey 等；基于上下文的标签，如 Golden Gate Bridge，2005-10-19 等；表示属性的标签，如资源发布者的姓名等；主观性的标签，如 funny，cool 等；组织性的标签，如 my paper，to-read 等。在此基础上，Xu 等认为高质量的

标签应具有如下特性：涵盖较多的方面；高使用度；尽可能少的数量；规范的格式；不包含特定分类的标签。Golder 等则更进一步地将标签分为 7 种类型[32]：标记资源的内容的、标记资源的类型的、标记资源的拥有者的、进一步描述分类的、标记资源的质量或特点的、自我引用的以及任务组织性的。基于这些工作，Bischoff 等[33]提出了一种启发式方法和机器学习方法相结合的标签类别识别方法，针对不同的类别的标签采用不同的识别方法，实现了自动化的标签分类。

标签间关系分析的一种方式是借助预定义的概念关系系统。Plangprasopchok 等[34]研究了用户对个人使用的标签所进行的分类组织过程，并给出了一种合并不同用户的标签组织以得到公共分类结构的方法。Angeletou 等[35]则利用 WordNet、在线本体等外部的形式化语义源来提供标签之间的关系，验证了这些形式化语义源在基于标签系统的搜索方面的作用。在文献[36]中，Angeletou 等进一步地对比了利用 WordNet 及在线本体的进行面向标签的查询扩展的效果，并指出 WordNet 可以扩展更多数量的标签，包含更多不同的语义，同时包含更多的上下位概念。而本体则在上层概念的扩充方面体现出了比较强的效能。Pan 等[37]利用本体来扩展大众分类法。这一方法的本质是将标签对应到本体上，并利用本体作为基础建立与其他标签间的关系。Tsui 等[13]则利用 WikiPedia 作为辅助分析标签间的层次关系的工具，通过利用标签所对应的 WikiPedia 网页中对标签词汇的分类提取标签间的层次关系。

预定义的概念关系系统虽然可以有效地帮助分析标签间的关系，但其通常只能用于有限的场景，并且无法有效地适应新出现的标签，因此一些工作也研究直接建立起标签间的关联。Solskinnsbakk 等[38]指出，缺乏清晰的结构是大众分类法固有的问题，并提出将用户对资源的一次标记作为事务对标签进行关联规则挖掘，通过关联规则置信度选取子节点构建标签层次关系的方法。Cattuto 等[15]研究了社会标记系统中标签间的语义距离，并对比了用于计算标签间相似度的 5 种方法。他们的研究指出，这些相似度计算方法均可以使用基

于共同标注关系的余弦距离来代替，并且共同标注关系向量和资源向量有助于发现标签的同义关系。Meo 等[39]提出了一种依据用户提交的查询标签进行扩展得到标签体系的方法。针对每一个查询标签，该方法计算其最相似标签，通过计算标签间的相似度得到候选标签集合，并利用标签出现的频率关系构建标签层次，及利用标签的覆盖率确定标签的语义粒度。Candan 等[40]提出了一种构建标签层次关系的方法以便提高资源导航的效率。根据标签标记资源的情况，该方法构建标签-资源关系矩阵，对该矩阵进行奇异值分解及奇异值削减，并针对削减后的矩阵应用余弦距离计算标签间的相似度。进一步的，该方法利用能量确定标签间的上下位关系，并利用有向图最小生成树算法构建标签层次关系。

1.1.3 基于标签的 Web 信息处理

标签系统开放、自由、灵活的使用特征令标签可以覆盖更加广阔的主题与领域，同时令标签携带了包括信息被发布和共享时的上下文[11]、信息本身不具备的特征[24]，以及信息多方面的属性[31]等大量潜在的信息。充分利用这些潜在的信息可以有效地辅助多种类型的 Web 信息处理任务，因此标签吸引了 Web 信息处理领域研究人员的大量关注。

标签系统的提出本身是作为对传统的基于受控词汇表、固定分类体系或本体的分类方法的替代。由于同样携带了关于如何依照概念对信息进行分类的信息，人们也研究利用标签系统的分类结果构建概念层次关系。Chen 等[41]研究了从标签结果中提取基本概念的问题。一个基本概念对应了一组概念中最具有区分度的概念，这样的一组概念很明显地适合作为受控词汇表使用。Monnin 等[42]提出基于 RDF 和标签行为模型的本体构建方法。它化解了各个标签模型之间的冲突，使得用户的每一次标记行为都可以被精确地描述，并通过具体化用户标记资源的动机分析潜在的标签语义。Tang Jie 等[43]首先通过概率主题模型对标签和文档进行建模，并计算标签之间的语

义差异程度。针对两个标签间上下位词、同义词和无关词的关系，定义了构建层次、合并以及保留3种基本操作，逐步完成标签本体的构建。Caro等[44]认为标签的层次体系应该体现出两方面的内容：标签的上下位关系以及标签之间的语义相似度。通过利用潜在语义分析标识标签的语义，以及利用标签在文章中出现的上下文，Caro等利用扩展布尔模型将标签逐步地组合到一个层次中，实现概念分类体系的构建。

对于单个的用户来说，标签是一种高度个性化信息组织和标注工具。因此用户个人的信息标注成果，从被用户标注的资源以及用户所使用的标签两个方面，体现出了用户自身的兴趣爱好。这种用户个人兴趣的直接体现使标签结果特别适合于进行个性化资源推荐任务。个性化资源推荐中一种重要的方法是协作性过滤方法。Meo等[45]提出了一个基于查询扩展和用户个性化信息强化的方法来提升协作性过滤系统的推荐性能。该方法为每一个用户维护一个个性化信息，利用PageRank方法计算标签的Rank，以此为基础返回一组可能的扩展标签供用户选择。Durão等[46]提出了一种混合考虑标签的文本相似度、语义相似度、热度、表现力以及标签和用户之间的关系等多种因素的基于标签的支持语义扩展的推荐系统，用于推荐相似的资源。Guan等[47]研究了基于图的子空间学习方法来实现基于标签的文档推荐方法。给定用户-标签、标签-文档、文档-用户及文档-文档4个关系矩阵，该方法学习一个用户-标签-文档的语义子空间，使其能最大限度地保存那些矩阵之间的关联结构。利用这一子空间，用户没有标记过的文档便可以推荐给用户。Yoshida等[48]利用标签扩展了基于内容的资源推荐。在他们的方法中，资源之间的相似度使用标签的评级来评价，与用户的日志文件中最为相似的资源则被推荐给用户。Gemmell等[49]对比了基本的资源推荐、基于标签的资源推荐以及一种基于线性混合的资源推荐方法，并表明混合方法可以在多种不同现实数据集上取得最佳的推荐质量。

作为对资源的一种分类、索引和查找工具，标签最为重要的应用是辅助用户对各种类型的信息进行搜索。Golub等[50]的研究结果

表明，虽然受控词汇表可以提供高质量的标签候选，但自由的标签却可以提供更多的新概念描述，以及超越受控词汇表所能提供的更多的信息。Abbasi 等[51]研究了用于标签搜索的查询松弛方法。该方法利用标签的分布情况确定标签的语义抽象层次，并利用资源-标签以及用户-标签矩阵来计算资源的上下文相似性，并以此为基础扩展查询线索，实现查询松弛。Abel 等[52]将资源的分组信息作为背景扩展标签信息，实现对信息的有效检索。该方法利用资源分组信息为原本不相关的信息提供一个共同的用户组作为上下文，利用一组标记规则如分组的标签等价于组内资源的标签等，实现标签信息的传递，并利用 FolkRank[17]方法实现结合上下文信息的资源排序方法。Amer-Yahia 等[53]认为，搜索过程不仅依赖于用户的需求与用户的历史行为，还依赖于来自其他用户所能提供的信息的帮助。基于这种观察，Amer-Yahia 等利用 LDA(Latent Dirichlet Allocation)或手工指定的主题向量发现用户的社区，并在这种社区的基础上实现了基于标签的资源搜索。Djuana[54]等从标签系统中提取领域本体，并反过来用于标签推荐。他们的研究结果表明，利用提取的领域本体来帮助对推荐结果进行重新排序，将可以进一步提升推荐的质量。

由于标签是由用户在使用资源的过程中逐渐添加的，其可以携带大量原始信息中没有明确指明的细节。这种额外的标注信息令标签可以揭示关于信息更多的内容，并进而帮助搜索系统提升搜索结果质量。Biancalana 等[55]提出了一种基于标签的查询扩展方法，该方法记录和处理用户的行为，以便依据用户的兴趣提供个性化的搜索结果。过程对用户完全透明，依赖用户进行的选择，提交的关键字以及点选的网页进行，并利用搜索过程获得的信息进行动态的更新。针对一个查询，该方法使用不同的扩展方法，以针对不同的语义项或领域进行扩展，并将结果按照不同的组别进行分组，呈现在一个页面中。Cattuto 等[15]的研究结果表明，标签的共现向量以及标签-资源向量有助于发现同义词的集合、拼写错误以及 WordNet 同义词集合，因此可以被用于查询扩展任务。Bhagwan 等[56]则探讨了使用标签改善桌面搜索质量的方法。桌面搜索面临的主要问题是冷启

动问题。针对这一问题，Bhagwan 等利用标签网站提供的标签-资源关系为桌面搜索提供基础标注，解决了冷启动问题，提供了一个有效的元数据推荐机制，同时保护了用户的隐私。

除了这些与分类及搜索直接相关的领域，标签系统也因其灵活性而在大量其他领域中发挥着特有的作用。Arabshian 等[57]利用基于标签的服务标注系统对服务进行索引，并通过将标签对应到一个本体系统上实现对 Web 服务的发现和选取。Gawinecki 等[16]也采用了类似的方法，通过结合标签系统的灵活性与结构化的 Web 服务功能和结构描述，实现了一个结构化的 Web 服务标签系统来提升 Web 服务的描述能力。Bouillet 等[58]也利用用户参与的服务标记实现对 Web 服务的建模。传统的 Web 服务模型如语义 Web 服务模型建模成本高昂，同时要求很多的参与、标注和管理。而 Bouillet 等提出的方法利用了标签系统不需要集中管理的优势，虽然描述能力低于传统的服务建模方法，但其已经能够提供基于标签的服务操作输入和输出描述，且在实验中体现出足够的标注能力。Chou 等[59]则利用标签来描述用户的服务需求和服务所能提供的功能，并使用形式化概念分析匹配用户的需求和服务的功能，为构建 MashUp 应用提供支持。除了 Web 服务相关研究之外，标签也在多媒体信息检索方面体现出了特有的价值。Abbasi 等[60]研究了结合用户信息从标签系统中提取具有特殊属性的图片的方法。通过将具有特殊性质的用户群体所使用的标签作为样本，以及将特定类型用户群体所使用的标签作为反例，该方法学习并识别具有特殊标记意义的标签，并通过这种方法从 Flickr. com 中识别地标建筑图片。

上述研究结果表明，标签及其组成的标签系统可以为大量不同类型的 Web 信息处理任务提供支持。这种广泛的应用，一方面证明了标签系统巨大的潜力，另一方面也对标签系统的标注质量提出了更高的要求。因此，如何进一步地提升标签系统的标注质量，以便促进标签系统在不同领域中的应用，使其发挥更大的价值，成为了标签领域中引起广泛关注的问题。