聚类搜索引擎研究进展综述①

周 鸿[1]，朱东华[2]，董萍萍[1]

[1](北京物资学院 信息学院，北京 1001149)
[2](北京理工大学 管理与经济学院，北京 100081)

摘 要：在综述国内外学者有关聚类搜索引擎和本体技术研究成果的基础上，试图梳理出现阶段该领域的研究热点和难点问题，为后续研究奠定一定的研究基础。分别从聚类搜索引擎的定义、研究现状，本体技术，基于本体的中文环境下语义聚类搜索等方面对已有的研究文献进行了系统的综述，并提出基于本体的聚类搜索引擎总体框架和成员引擎的调度策略；在上面基础上提出对未来研究的展望。

关键词：聚类搜索引擎; 本体；语义搜索；调度策略

Survey of Research Progress on Clustering Search Engine

ZHOU Hong[1], ZHU Dong-Hua[2], DONG Ping-Ping[1]

[1](School of Information Science & Technology, Beijing Wuzi University, Beijing 100049, China)
[2](School of Management & Economics, Beijing Institute of Technology, Beijing 100081, China)

Abstract: Based on summarizing the research results of clustering search engine and ontology technology by scholars at home and abroad, the efforts are made to attempt to sort of research hot issues and difficult problems in this field to pave certain research foundation for the follow-up research in the future. Through respectively summarization of the owned research documents about definition of clustering search engine, status quo of research, ontology technology as well as semantic clustering search in the context of Chinese based on ontology, the overall framework for clustering search engine dispatching strategies for member engine based on ontology are proposed, and the expectation for future research based on aforementioned result is proposed.

Key words: clustering search engine; ontology; semantic search; scheduling strategy

在信息资源管理领域，Internet 上的信息资源具有海量、分布、动态、复杂、开放等特点，用户如何从这些海量的数据中查找自己所需要的有用的信息，为国家的科研和社会应用服务提供信息基础，这就需要一种新的技术，能自动地从 Web 上发现、抽取和过滤信息同时满足用户在特殊的中文环境下的语义、语用和知识上的需求。

在聚类搜索引擎的理论研究中，O.Zamir[1]提出 Web 信息聚类的 STC(Shared Term Clustering)方法，该方法的主要缺点有：仅仅依靠共同词组进行聚类，没有考虑语义的问题，其中的关键词组发现算法没有考虑词组的稳定性和完整性，而且不能适用于中文等语言，直接以后缀树的层次结构作为类的层次结构，并不合理，D.Cutting 等人[2]提出 Web 信息聚类的 Scatter/Gather 系统，由于是采用传统的启发式聚类算法进行聚类，难以避免启发式聚类算法的种种缺点。Y.Wang[3]等人提出基于超链接(hyperlink)进行万维网信息聚类，这种方法需要下载并分析实际的网页，因此不可能做到在线聚类。在目前国内外针对聚类搜索引擎的研究过程中，存在以下瓶颈和问题：

(1) 聚类引擎标签的可读性差。目前基于“关键词查询+用户自行浏览”的搜索引擎信息交互方式造成了用户信息需求的传递瓶颈，聚类搜索引擎的类标签是为了解决这一个问题。但是在目前的研究中，因其普

① 基金项目:北京市教委科研基地建设项目(WYJD200902);北京市属高校人才强教项目(PHR200906210)
收稿时间:2011-08-12;收到修改稿时间:2011-09-11

遍是基于在线聚类的方法，导致类标签的可读性、导向性以及和用户查询的相关性及差，如何提高类标签可读性、导向性、主题相关性三个指标的质量是目前急需解决的首要问题；

(2) 聚类搜索引擎的检索效率。在线聚类的传统方法，具有服务器系统负载轻、速度快的优点，但是面临在线聚类的标签可读性及差的问题；线下聚类，又面临引擎服务器承担的负载难以承受，其检索效率难以提高；

(3) 聚类搜索引擎的语义问题。结合中文搜索环境，聚类引擎的理论研究还没有克服搜索结果页面和标签之间的中文语义问题，应引入本体技术来解决上述两个问题。

本体是用于描述或表达某一领域知识的一组概念或术语，它可以用来组织知识库较高层次的知识抽象，也可以用来描述特定领域的知识[4]，本体技术已经发展成为知识表示、知识管理、知识共享、知识复用的主流技术之一，同时本体是 Web 信息在语义层次上共享和交换的基础，可以大大加强 Web 的功能，其正成为自然语言处理、Web 信息检索、数据库和知识库的管理、异构数据集成、数字图书馆、GIS、语义 Web 等研究领域共同关心的一个核心问题[5]，同时常规关键词的信息检索技术已不能有效揭示海量信息资源和满足用户在语义、语用上和知识上的需求，在越来越多的研究中开始集中在利用本体解决中文环境下的语义搜索和标签聚类问题。

1 聚类搜索引擎

针对目前“关键词查询+用户自行浏览”的信息交互方式造成了用户信息需求的传递瓶颈，越来越多的研究集中在聚类搜索有关的研究方面；聚类搜索引擎是元搜索引擎与 Web 聚类技术的整合，它通过元搜索引擎获取来自不同传统搜索引擎的搜索结果，然后对搜索结果进行聚类，形成簇集，每个簇中均包含对应的文档集，聚类搜索引擎与传统搜索引擎的最大区别就在于搜索结果表现形式的不同。传统搜索引擎的搜索结果是一按相关性排序的结果列表，而聚类搜索引擎则是将搜索结果进行再聚类，形成类标签和对应的类内容的过程。其代表有 Vivisimo、Carrot2、Mooter、Wisenut、webmenu、bbmao 等。

1.1 聚类与元搜索引擎

随着元搜索引擎技术的成熟，聚类搜索引擎的基础架构包括数据获取的来源、数据的清洗、和类标签的融合等都有充足的理论基础。元搜索引擎就是对多个独立搜索引擎的整合、调用、控制和优化利用，Meta-Search 没有自己的索引数据库,如 MetaCrawler。

考虑到元搜索引擎的特殊性，聚类搜索引擎在数据和类标签的融合过程中，可以充分利用成员元搜索引擎提供的结果基本信息，如标题、摘要、连接等相关数据描述。而目前几乎所有基于传统机器人搜索引擎都提供标题和摘要等描述，对于数据大小、快照等不全提供[6]。

聚类搜索引擎充分利用现有搜索引擎的搜索结果，快速、有效地将信息进行整合,提供给用户更加有价值的知识，结合在前面提到的元搜索引擎和目前的聚类搜索引擎的瓶颈，聚类搜索引擎应是基于传统机器人全文搜索引擎和元搜索引擎基础上的搜索引擎。

聚类搜索引擎基于元搜索，并不属于元搜索引擎，因其有其基于摘要的索引库、关键词标签库等；聚类搜索引擎是在线聚类和缓存索引机制相结合、动态演化、基于语义、可视化和摘要描述的检索引擎。

1.2 聚类搜索引擎的研究现状

聚类搜索引擎一般是元搜索引擎(Meta Search Engine)与 Web 聚类技术的整合，通过元搜索引擎获取来自不同搜索引擎的搜索结果，然后对搜索结果一般是文档的一个片段，而不是整篇文章)进行聚类，形成类标签和类结果，每个类标签中均包含对应的类内容。

在聚类搜索引擎的应用研究中，D.Cutting[2]等提出的称为“Scatter/Gather”的技术，试图更合理的组织搜索结果，Scatter/Gather 是较早对搜索结果进行聚类的系统，其基于 Buckshot 和 Fractionation 两种聚类算法基础。

Agglomerative Hierarchical Clustering(AHC)是基于 K-means，Single-pass 等基于距离的聚类算法，但精度很低，类标签很难表示。

Grouper 通过 HuskySearch 元搜索引擎获取搜索结果并利用基于短语(phrase)的后缀树聚类(Suffix Tree Clustering)算法对其动态聚类成带有类标签的簇集。

Carrot2 和 Semantic Hierarchical Online Clustering 两者都用后缀数组来发现关键短语，不同的是前者使用 SVD 算法来得到类标签，再利用向量空间模型形成

簇集，而后者利用 SVD 来得到类标签和形成簇集。

Vivisimo 基于的原理是一种叫做准确描述所有配对(concise all pairs profiling)(简称为 CAPP)的方法，这种方法着眼于形成可描述的聚类.它的基本原理是将所有的类别成对的进行比较，找出能够将每一对类别区分开来的特征，然后对那些特征进行组织，形成最后的描述，保证每一对至少有一个特征能够将它和其他对区别出来，Vivisimo 自动聚类所依据的是搜索引擎返回的网址、标题和简单描述，而不是整个网页。

2 基于本体的聚类搜索引擎

聚类搜索引擎众多的研究多集中在对查询和页面文档的相关度以及页面聚类的定性描述和研究上，而聚类引擎需要更多地专门从类标签聚类、用户聚类的角度出发，突破搜索引擎目前存在的瓶颈，解决存在的不足。本体是语义检索的重要辅助手段，由于本体本身具有一定的推理能力，可以利用本体进行查询扩充，从而使检索的结果更加全面，假定本体库已经建立完备，可以考虑使用类和属性的继承关系来对查询进行扩展，以期提高查全率和查准率。

2.1 本体技术

Studer[7]等对多种本体定义进行深入研究，认为本体是“共享概念模型的明确形式化规范说明”，这个定义包含 4 层含义：概念化 Conceptualization)、明确化(Explicit)、形式化(Formal)和共享(Share)。

定义 1 本体概念化定义为:O=(D,W,R)其中 D 是一个领域，W 是该领域中相关的事务状态的集合，R 是领域空间<D,W>上的概念关系的集合，本体论是采用某种语言对概念化的描述。目前，普遍认为本体包括以下 5 种元素:类、关系、函数、公理和实例。已有的本体构建工具中较成熟的有 Ontolingua、webOnto、OntoEdit 等。

在国外的研究过程中，有几个比较典型的实验系统，如 KEUOA 系统，这是一种通过使用一种简单的用户定义的知识抽取模式来从互联网页上抽取知识结构的工具；Artequakt 的项目利用一个基于 ontology 的知识抽取工具来实现连续的知识支持和引导信息抽取，能够搜索在线的文档，并且把其中符合事先定义好的结构的知识抽取出来；OFEE（Ontology-based Fuzzy Event Extraction）系统，是一个基于 ontology 的汉语新闻摘要的模糊事件抽取代理系统。

2.2 基于本体的中文语义搜索

本体和 web 技术结合是近年来国际上 Web 智能等领域的重要研究方向，其应用领域日益扩大。目前，人们已进行了许多有关语义 Web 基础架构如本体语言 OWL、编辑器、推理引擎等方面的工作，然而，面对快速增长的 Web 信息，很多基于 Web 的应用面临着相关领域本体缺乏的问题，因为许多的本体构建严重依赖于以专家为中心的方式实现的，这种以手工为主的构建不仅代价很高，无法进行大规模扩展，同时要促使大量的用户和领域专家为语义 Web 来构建本体也存在相当的困难，因此研究自动的，通用的领域本体构建方法是解决这一问题的关键。

从现有知识源(如文本、词典、遗留知识库或本体、数据库模式等)获取领域知识、以(半)自动方式构造或改编本体，即所谓的本体学习(Ontology Learning)，是开发本体的有效途径[8-10]。

国外的研究中，Farrar[11]把语义 web 语言学和本体结合起来。Yuli[12]对 image 的索引、分类、萃取等应用本体进行了研究。Jerry[13]根据可靠性语义的计算提出了本体的匹配和评估算法。Qing[14]对 owl 语言的构建工具在论文中作了大量工作。Yongchun[15]针对牛奶工厂应用 web 本体语言进行了建模和应用。Ho Wang[16]针对不确定信息领域本体模糊匹配的算法进行了研究和扩展。

与国外相比，国内在领域概念的自动抽取方面，特别是中文领域概念的自动抽取的研究工作相对较少。程勇[17]基于本体的不确定性对知识管理进行相关研究。上海交通大学的杜波等人[18]提出了一种将统计方法与规则方法相结合的专业领域术语抽取算法。山西大学郑家恒等人[19]提出采用非线性函数与“成对比较法”相结合的方法，综合考虑位置和词频两个因素，给出候选词的权重，实现了关键词的自动抽取。东北大学的陈文亮等人[20]提出利用 Bootstrapping 的机器学习技术，从大规模无标注真实语料中自动获取领域词汇。浙江大学的刘柏嵩等提出一种 Web 页面中自动抽取本体 WebOntLearn 的方法[9,21]，从 Web 页面数据中找出本体语义概念的模式及其关系，并通过分析同一应用领域 Web 页面集来半自动化地抽取 Web 本体。

2.3 基于本体的 web 聚类

在国内的研究中，邓健爽等人（2007）[22]研究了比较简单的基于搜索引擎的关键词自动聚类法。史庆伟等

人结合STC算法和变色龙算法提出了一种中文网页的层次聚类方法-STCC 算法。黄德才等人[23]研究了基于主题相似度模型的 TS-PageRank 算法。陈再良等[24]人提出了一种改进的分布式 dPageRank 算法。

国内 bbmao[25]则对搜索引擎的检索结果进行分类，在线对标签进行分类，和国外的 carrot2、viviismo 等引擎功能类似，但聚类树比较简单，同时在语义检索功能在满足用户需求方面相对较弱。

武汉大学的余传明[26]对基于本体的语义检索的处理进行初步的研究。徐敏[27]研究了基于关联规则的在线 web 检索等相关理论。国防科大的宋俊锋[28]结合语义 web 的领域本体的表示、推理、集成进行了相应的研究工作。赵永屹[29]介绍了搜索引擎快速聚类系统的初步的原理性的设计与实现。刘群[30]等基于知网的语义相似度的计算的工作进行了相应的研究。施水才[31]等在元搜索引擎基础上，介绍当前主要的聚类算法：K 均值划分法和层次凝聚聚类法，并在此基础上提出基于元搜索结果将两种聚类算法相结合的聚类方法。周登朋[32]针对搜索引擎的结果的一些聚类算法设计了DIRS(Document Information Retrieval System)系统。北京大学的余晋[33]等利用 STC 和 SWC 算法提出了PinkySearch 系统。

在国外的研究中，Collection Building 项目，简称 CBP 项目，是美国国家科学数字图书馆支持下的一个子项目，在该项目中，它将馆藏定义为关于某个主题的 web 网页、PDF、ps 等文件组成的集合，它通过主题爬行来逐渐生成。Focus Project 是印度学者 S.Chakrabarti[34]在伯克利大学计算机系读博士期间所从事的一个项目，它对 web 资源主题的定义既不是采用关键词也不是加权词矢量，而是一组具有相同主题的网页。

O.Zamir 和 O.Etzioni[1]采用后缀树（Suffix Tree）数据结构给出了一种网页快速聚类的方法，称为 STC(Suffix Tree Clustering)算法，并开发了 Grouper，用来对(元)搜索引擎 HuskySearch 的结果进行聚类。

David Weiss[35]在他 01 年的硕士论文中详细描述了他把 STC 用于波兰语的 Carrot System，由于波兰语的语法和构词与英语有很大区别，David 建立了一个波兰语词库，并且开发了一个 quasi-stemmer 对波兰语进行词干截取。

D.Zhang[36]提出 SHOC 算法，并指出这是 STC 的一个延伸，不仅适用于英语，并且适用于东方语种，如中文，其两个新颖之处在于：关键词组的发现和正交聚类，关键词组的发现是基于这样三个标准：完整性、稳定性、重要性。

2.4 基于本体的聚类搜索引擎总体框架

基于本体的语义检索模型结构如图 1 所示：该结构包括人机交互模块、查询优化和本体标签聚类模块、成员引擎调度模块、web 文档聚类等几个模块组成。人机交互层包括用户、查询界面、查询请求是用户和系统通讯的接口。

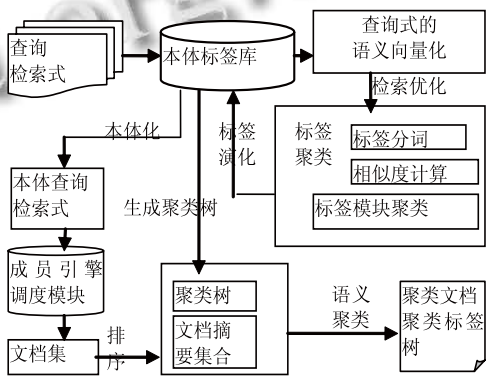

图 1 基于本体标签的信息检索模型

查询优化主要涉及对用户查询检索式的优化。类标签聚类模块负责聚类搜索引擎的标签原始数据的获取、分词、关键完整性判断、语义相似度计算、标签本体模块化、聚类树的生成；最后通过类标签加载相应的文档摘要索引以可视化的聚类树提交给用户。

成员引擎调度模块是页面采集的调度控制核心，具体的成员引擎数据库为：baidu、Yahoo、sougou、zhongsou、msn、天网、yisou、InfoSeek、AltaVista、Lycos、Webcrawler；页面的采集和一般的元搜索引擎的采集过程类似，数据都来自于成员引擎数据库的全文检索的基础上；和一般的 metasearch 不同的是该调度模块结合标签库进行查询和标签树的拟合、计算查询和成员引擎数据库被索引文档的相关性。解决了传统元搜索无法对成员引擎数据库定性定量评价的瓶颈，根据相似度选择成员引擎，尽量避免在聚类搜索引擎的查询过程将查询请求不加区分地全部提交给成员数据库执行。文档聚类主要涉及基于本体的后缀树聚类算法，对文档进行语义聚类、用类标签进行索引标记，聚类引擎索引库基于文档摘要进行检索，拥有

摘要索引库。

3 聚类搜索引擎的调度策略

聚类搜索引擎的调度策略是研究聚类搜索引擎如何为用户选择数量合适并且和类标签相关度最高的成员引擎数据库，以较小的资源消耗，来获得最大的搜索收益；

定义 2 成员搜索引擎数据库 D 在类标签 L_i 中的被引用摘要 $R'(l_i,D)$ 由两部分组成：

(1) $D'(l_i)$ 的实际被引用总数，$|D'(l_i)|$；

(2) $D'(l_i)$ 中包含的被引用的文档权重；

假设从 m 个成员搜索引擎数据库 $D_1,D_2,\ldots,D_m$，D_i 的被引用摘要为 $R'(D_i)=\{R'(l_1,D_i),R'(l_2,D_i),\ldots,R'(l_n,D_i)\}$，其中 $R'(l_1,D_i)$ 表示数据库 D_i 在类标签 l_j 的被引用摘要。

则成员引擎的数据库的被引用摘要 $R'(D_i)$ 对于类标签 L 的相似度，计算公式如下：

$$sim(L,R'(D_i))=\sum_{j=1}^{t} lw_{ij} * disl_j$$

其中 lw_{ij} 表示第 i 个数据库中被类标签 l_j 引用文档的权值之和，即 $lw_{ij}=\sum p'(doc_j|D_i(l))$；其中 $disl_j=\frac{1}{t}\sum_{i=0}^{t}(lr_{ij}-alr_j)$，$lr_{ij}$ 表示标签 l_j 在 D_i 和所有引擎数据库中被索引权重的比例，即 $lr_{ij}=\frac{lw_{ij}}{\sum_{i=1}^{t} lw_{ij}}$，$alr_j$ 表示所有 lr_{ij} 的平均值。

在聚类搜索引擎的调度策略过程，首先计算各成员引擎的数据库与用户兴趣类别标签 L 的相关度，根据相关度对成员引擎进行排序，选择排名最前的几个成员搜索引擎数据库为用户提供查询服务。

4 聚类搜索引擎研究展望

因为中文语言环境的复杂性，搜索引擎在长期以来，语义检索理论的发展比较低迷，研究的切入点应落实在结合本体技术上来解决这一问题，对文档和类标签进行语义聚类，聚类搜索引擎的理论研究还有大量的工作需要完成，进一步的研究工作主要基于以下几个方面展开：

(1) 本体标签库原型的构建工作；综述中提出的解决方案是通过本体技术，根据语义来自动构建类标签和文档索引库。本体库原型可以在知网的基础上，来进行研究和实现，假设知网的义原相似度能较好的表达中文的语义需求基础上的。

(2) 人机交互界面；聚类搜索引擎的进一步研究，将着重研究聚类树和数据可视化等一些理论方法。

(3) 聚类搜索引擎正处于方兴未艾的时期，对聚类搜索引擎的评价目前的研究更加少，包括目前普通搜索引擎的现有评价体系也非常不足，其都是在搜索引擎之间总体横向的评价对比的基础上，对于搜索引擎内部的算法以及索引机制以及人机交互界面建立一个综合评价指标体系也是需要解决的课题。

(4) 缓存索引机制；对于在线聚类的数据，元搜索引擎是没有索引库的，聚类引擎建立缓存索引机制，在对聚类数据进行缓存索引机制的同时，相应的对类标签库进行交互融合；

(5) 面向搜索用户的个性化聚类引擎；如何基于对用户宏观行为和历史行为的分析，让聚类引擎系统加深对用户信息需求的理解，进而缓解信息需求的传递瓶颈，来提高个性化搜索的水平；这一方面应关注用户使用数据的挖掘、聚类等理论。

(6) 对聚类搜索引擎的 API 接口研究，提供专业化搜索；专业性的搜索引擎为专门收录某一行业、某一主题和某一地区的信息而建立，具有很高的实用价值。聚类搜索引擎融合了能涵盖大多数行业和领域的元成员搜索引擎，因此聚类引擎是否可以结合本体语义技术提供 API 为科学研究和社会，为特定的行业和主题建立专业化的领域搜索需要进一步的研究和探索。

参考文献

1 Zamir O. Clustering Web document: A phrase-based method for grouping search engine results. University of Washington, 2000.

2 Cutting D, Karger D, Pedersen J, Tukey JW. Scatter/Gather: A cluster-based approach to browsing large document collections. Proc. of the 15th Annual International ACM/SIGIR Conference, Copenhagen, 1992.

3 Wang Y, Kitsuregawa M. Link-based clustering of Web search results. Proc. of the Second International Conference on Web-Age Information Management (WAIM2001). Xi'An,

China, Springer−Verlag LNCS, 2001.
4 Wiliam S, Austin T. Ontologies. IEEE Intelligent Systems, 1999 Jan/Feb:18−19.
5 付相君.基于本体和 Semantic Web 技术的产品知识集成基础研究[博士学位论文].杭州:浙江大学,2005.
6 刘畅等.元搜索引擎的调查分析.现代图书情报技术,2004,114(9):40−43.
7 Studer R, Benjamins VR, Fensel D. Knowledge engineering, principlesand methods. Dataand Knowledge Engineering, 1998,(25):161−197.
8 杜小勇,李曼,王珊.本体学习研究综述.软件学报,2006,17(9):1837−1847.
9 刘柏嵩.基于知识的语义网:概念、技术及挑战.中国图书馆学报,2003(2).
10 Sabou M. Leaming Web Service Ontologies: an Automatic Extraction Methodand Its Evaluation. ISWC 2005.
11 Farrar. An ontology for linguistics on the Semantic Web. The University of Arizona, 2003.
12 Yu L. Ontology-based large-scale image classification,indexing and exploration. The University of North Carolina at Charlotte, 2007.
13 Jerry. Mapping local ontologies:Authentic semantics for learning object evaluation. Simon Fraser University (Canada), 2006.
14 Qing. OntoKBEval: A support tool for OWL ontology evaluation. Concordia University (Canada). 2006.
15 Yong C. The application of Web ontology language for information sharing in the dairy industry. McGill University (Canada). 2006.
16 Wang H. An extension of the crisp ontology for uncertain information modeling-fuzzy ontology map. Hong Kong Polytechnic University (Hong Kong), 2007.
17 程勇.基于本体的不确定性知识管理研究.中国科学院,2005.
18 Du B, Tian HF, Wang L, Lu RZ. Design of domain-specific term extractor based on multi-strategy. Computer Engineering, 2005,31(14):159−160.
19 Zheng JH, Lu JL. Study of an improved keywords distillation method. Computer Engineering, 2005,31(18):194−196.
20 Chen WL, Zhu JB, Yao TS. Automatic leanring field words by bootstrapping. Proc. of the JSCL Beijing: Tsinghua University Press, 2003. 67−72.
21 刘柏嵩.基于本体的知识管理关键技术研究.情报学报,2005,(1).
22 邓健爽,等.基于搜索引擎的关键词自动聚类法.计算机科学,2007.
23 黄德才,戚华春,钱能.基于主题相似度模型的 TS-PageRank 算法.小型微型计算机系统,2007,(3).
24 陈再良,凌力,周强.dPageRank—一种改进的分布式 PageRank 算法.计算机应用,2006,(1).
25 http://www.bbmao.com.
26 余传明.基于本体的语义信息系统研究[博士学位论文].武汉:武汉大学,2005.
27 徐敏.基于数据挖掘的 Web 信息检索研究[博士学位论文].南京:南京航空航天大学,2006.
28 宋俊锋.面向语义 web 的领域本体表示、推理、集成及其应用研究[博士学位论文].长沙:国防科技大学,2006.
29 赵永屹.中文专业搜索引擎检索结果聚类的研究与实现.北京:北京理工大学,2006.
30 刘群等.基于《知网》的词汇语义相似度计算.中文计算语言学,2002.
31 施水才,等.基于元搜索的聚类挖掘引擎.情报检索,2007.
32 周登朋.搜索引擎搜索结果的聚类研究[硕士学位论文].上海:上海交通大学,2007.
33 余晋,等.pinkySearch:基于聚类的元搜索引擎.计算机科学, 2005.
34 Chakrabarti S. Automatic Resoruce Compilation by Analyzing Hyperlink Structure and Associated Text. Proc. of the International WWW Conference. Brisbane, Australia, 1998.
35 Weiss D. A Clustering Interface for Web Search Results in Polish and English [MS Thesis]. Poznan University of Technology. 2001.
36 Zhang D, Dong Y. Semantic, Hierarchical, Online Clustering of Web SearchResults. Proc. of the 6th Asia Pacific Web Conference (APWEB). Hangzhou, China. Apr 2004.