基于 WordNet 和 SUMO 本体集成的自动语义检索及可视化模型

Automatic Semantic Retrieval and Visualization Model Based on WordNet and SUMO Ontology Integration

胡泽文

Hu Zewen

摘　要　针对语义检索在实际应用中面临的用户查询意图获取困难、潜在语义索引计算复杂、领域本体覆盖范围小、概念语义类型不丰富、自动化程度低等问题，提出基于 WordNet 和 SUMO 本体集成的自动语义检索及可视化模型。实验表明这种模型能够过滤掉大量与用户查询无关的信息，提高信息检索系统的检准率，并很好地满足用户可视化和个性化检索需求。图 6。表 2。参考文献 34。

关键词　本体集成　语义检索　可视化　概念语义图　模型

Abstract: There still exists some problems in the practical application of semantic retrieval, such as hardness to acquire user's query intent, complex computing process of latent semantic index, small area of domain ontology coverage, poor concept semantic types, lower automatic level, etc. Aiming at these problems, this paper puts forward the automatic semantic retrieval and visualization model based on WordNet and SUMO ontology integration. Experiment shows: the proposed model can filter out a large number of information unrelated to user query subject, and not only improve the precision of information retrieval system, but also well satisfy user's retrieval demand of visualization and individuation.

Keywords: Ontology Integration; Semantic Retrieval; Visualization; Concept Semantic Map; Model

1　国内外研究现状

1.1　本体集成

本体在实现语义 Web 服务、语义分类与智能检索方面有着举足轻重的地位，近年来成为国内外学者关注的焦点，一些著名机构、企事业单位设计了很多语义丰富且具有实用价值的本体库，如 SUMO[1]、WordNet[2]、DBpedia[3]、OpenCyc[4]、HowNet[5] 和企业领域本体库[6] 等，由于开发者、应用目的和应用领域有所不同，它们在本体描述语言、内容侧重点、存储方式、查询语言、构建平台等方面存在很大差异，导致不同本体库之间无法进行有效通信和互操作，阻碍了本体在语义 Web、智能检索、文档语义分类等领域中的广泛应用。本体集成是能够有效解决本体异构问题，推动本体在语义 Web 服务、智能信息检索、文档语义分类等领域广泛应用与深度发展的最佳途径[7]。

目前国外学者和机构对本体集成领域的研究已有 10 多年的历史，设计了一批功能丰富、技术先进、可操作性强的本体集成工具，如 PROMPT[8]、OntoMerge[9]、GLUE[10]、OntoMap[11]、COMA + +[12]。并针对本体集成中存在的问题，提出很多不同的本体集成解决方案，其中应用较为广泛的本体集成解决方案是德国卡尔斯鲁厄大学 AIFB 研究所 Stumme G 和 Maedehe A 在 2001 提出的基于形式概念分析（FCA）的本体集成方法[13]。该方法能够有效解决本体集成领域中存在的一些问题，不过通过形式概念分析构建概念格的过程较复杂，因此仅适用于轻量级本体的集成。与 FCA 解决方案不同的是美国学者 Hitzler P、Krotzsch M 和 Ehrig M 等人在 2005 年提出的基于范畴论的本体集成方法[14]。由于范畴论具有非常好的数学理论基础，因此该方法适合处理重量级本体的集成问题，不过目前该方法仅处于探索阶段，实现起来还比较复杂。

国内本体集成研究起步较晚，最早起源于2003年，复旦大学张凯等学者在《计算机工程》期刊上发表题为“基于本体集成的资源共享平台”的文章，至今国内学者已发表了23篇标题含有“本体集成”字样的文章（检索数据库：万方期刊数据库；检索时间：2012－2－21）。其中，中国国防科技信息中心卢胜军等提出一种基于WCONS的本体集成方法[15]，燕山大学张忠平提出基于OWL DL图闭包和RDFS图闭包的本体集成方法[16－17]。这些方法能够有效推动本体集成领域的发展，不足的是，第一种方法主要论述了本体集成的方法流程及本体集成的一些操作，而对本体集成的核心部分“概念相似度计算及本体概念语义映射关系的建立方面”论述较少；第二种集成方法需要定义大量规则，集成过程比较复杂。另外，国内学者和机构在本体集成实践领域的研究成果过少，目前国内比较权威的本体集成工具有南京大学瞿裕忠和胡伟等人研发的一款本体匹配工具Falcon-AO[18]和天津大学魏哲雄等人研发的一款本体合并工具雏形OnMerge[19]。

综上所述，目前国内外学者在本体集成领域的研究主要集中在通用的本体集成模型、框架和方法，集成的层次也主要集中在概念层次，这些集成方法主要应用于中小领域本体的集成。而对于大型本体库，由于侧重点不同，集成起来非常耗时耗力。笔者认为，如果在具体领域应用到这些本体库的不同部分时，只需对这些本体库中的不同部分建立映射关系，然后基于映射关系对需要用到的那部分本体实体进行集成即可。

1.2 语义检索

网络信息资源的海量无序、难以有效利用与人们日益增长的信息需求之间的矛盾，成为目前信息化社会急需解决的难题。传统基于关键词匹配的信息检索方式虽然能够在一定程度上缓解这一矛盾，但由于其不能完全表征文档和查询语句中蕴含的语义，造成文档的误检和漏检；过分依赖用户的检索式，缺乏语义分析能力和语义扩展能力，难以保证较好的查准率和查全率，无法有效满足新一代语义Web环境下用户对海量网络信息资源语义分类、语义导航与语义检索的需求问题。对此，国内外学者提出一系列语义检索方法。

澳大利亚Park L. A. F和Ramamohanarao K提出基于概率潜在语义分析的信息检索方法[20]。但概率潜在语义索引计算复杂度高，存储空间要求高，在大规模数据集上进行潜在语义索引时，需要计算机具备强大的处理能力和方法。Mehul B和Wenny R等学者提出本体驱动的语义检索方法[21]。但该方法没有一个有效的查询和内容语义分析、处理和匹配的模式供计算机自动处理。

国内学者在信息检索新理论和新技术研究方面紧跟国外最新研究动态，针对传统信息检索方式所带来的问题，提出了一系列解决方法。最具代表性的是上海大学周文等提出的一种基于潜在语义分析的智能信息检索方法[22]，该方法能够有效解决传统信息检索方式缺乏语义支持造成的检索精度低的问题，不过文章没对用户查询进行语义扩展，可能造成检索结果的不全面，并且没有解决文本和查询中存在的同义词、多义词对文本检索精度的影响。何琳、候汉清等提出一种基于领域本体的语义检索方法，该方法相比传统的关键词检索可以发现潜在的、隐含的语义结果，具有较高的检准率和检全率[23]。美中不足的是该方法在语义关系判断上，对于本体库中存在的语义模式可以成功匹配到，但对于无法精确匹配到的语义类型目前还不能做到模糊匹配，并且领域本体的覆盖范围、概念及其语义关系的丰富程度会严重影响检索性能。

限于技术实现方面的难题和有效方法的缺失，目前国内外学者和机构对资源语义分类与检索技术在具体领域实践方面的研究还不够深入，大部分还停留在可行性分析与实验验证阶段，具有实用价值的语义分类与检索系统还较少。

2 基于WordNet和SUMO本体集成的自动语义检索及可视化模型

针对目前本体集成领域中提出的集成方法过于复杂，没有针对具体领域具体应用的集成方案和信息检索领域中存在的问题，笔者将本体集成和信息检索领域中的一些理念、技术和方法融

合起来，引入词汇覆盖率较高的英文同义词典 WordNet 和概念及语义关系丰富，涵盖领域广泛的标准上层本体 SUMO，并借助两者之间的映射关系，设计和实现一个基于 WordNet 与 SUMO 本体集成的自动语义检索及可视化模型，如图 1 所示。该模型分为用户查询与结果反馈、本体集成、融合与可视化、检索引擎四个模块。模型实现的基本流程：（1）基于 WordNet 同义词集与 SUMO 本体概念之间的映射文件，编写正则表达式，抽取 WordNet 同义词集及对应的 SUMO 本体概念，形成涵盖 WordNet 同义词集和 SUMO 本体概念之间一一映射关系的同义词集 - 概念两层索引库，然后以同义词集 - 概念两层索引库中的概念为基本索引单元，利用大型本体切割工具从 SUMO 大型本体文档中切割出概念对应的大小适度的本体片段，形成同义词集 - 概念 - 本体片段三层索引库。（2）对用户查询语句进行自动分词、提取关键词。然后利用 Mysql 查询语言从建立的同义词集 - 概念 - 本体片段三层索引库匹配出用户查询关键词对应的同义词集和概念，并在用户信息检索界面上将概念对应的本体片段可视化成本体概念语义图。（3）系统会自动将查询关键词及其对应的同义词集、概念及其上下位概念、同义概念组合成语义查询向量，检索信息资源库，并将检索结果按时间或相关度排序后反馈给用户。如果用户对查询结果不满意，可以通过可视化的本体概念语义图选择与需求主题更相关、更专业的领域概念，系统会根据用户选择的领域概念，将此概念的上下位概念、同义概念和对应的同义词集组合成新的语义查询向量进行信息资源的检索。

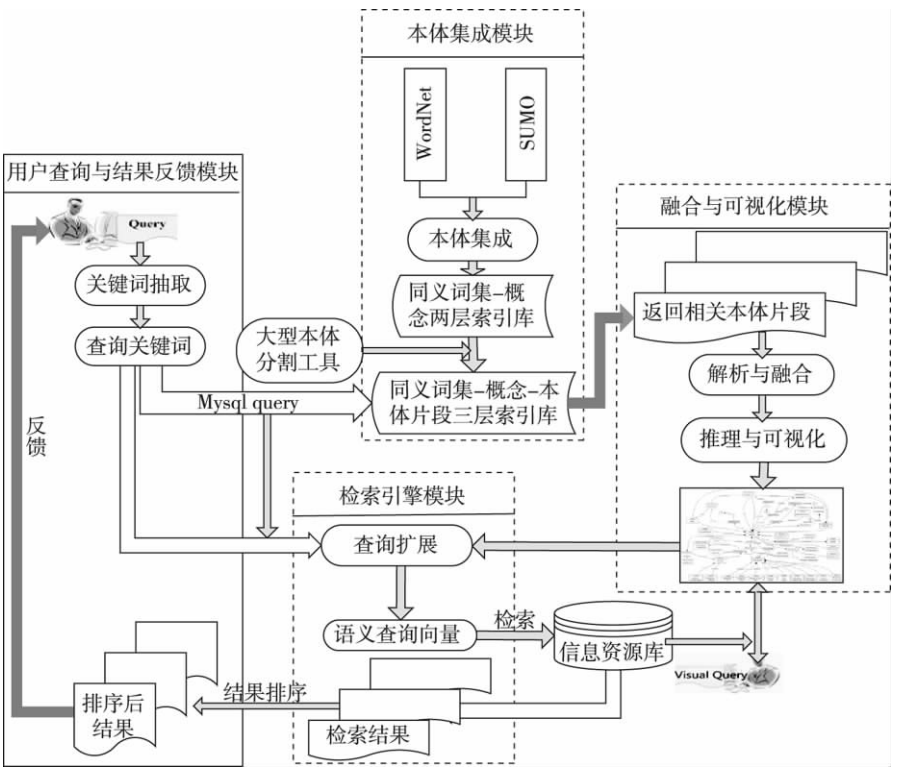

图 1　基于 WordNet 和 SUMO 本体集成的自动语义检索及可视化模型

2.1 本体集成模块

该模块输出的同义词集－概念－本体片段三层索引库是模型能够成功运行的基础和核心，不仅可以提高模型的性能和自动化水平，还可以通过本体片段的可视化增强用户检索体验。对于它的实现，主要分为两部分：（1）基于 WordNet 同义词集和 SUMO 本体概念之间的映射关系构建同义词集－概念两层索引库；（2）以同义词集－概念索引库中的概念为基本处理单元，利用本体切割工具从大型 SUMO 本体文档中分离出以该概念为中心单元的本体片段，存储成易于可视化的文档格式，并在同义词集－概念索引库中新增本体片段的路径字段，形成同义词集－概念－本体片段三层索引库。

2.1.1 同义词集－概念两层索引库构建

SUMO 本体库是 IEEE 标准上层知识本体工作小组为了发展标准的上层知识本体 SUO（Standard Upper Ontology）而设计的，主要专注于领域概念，基本上涵盖所有领域的概念实体及其之间的语义关系，可以用于检索、语言学和推理等方面的研究和应用[24]。WordNet 是美国普林斯顿大学认知科学实验室心理学家乔治 A. 米勒等学者在国家自然科学基金项目的支持下，为解决词汇之间缺乏语义关联，查找利用起来困难等问题，将具有相同意义的词汇组合成同义词集，根据同义词集之间的语义关系，利用语义网络将其组织起来[25]。WordNet 同义词集与 SUMO 本体概念之间存在着映射的关系，WordNet 同义词集能够全部映射到 SUMO 本体中的相应概念[26]。有关 WordNet 同义词典与 SUMO 本体的具体概况及两者之间映射的详细机制可以参考笔者在《现代图书情报技术》上发表的专题文章“WordNet 与 SUMO 本体之间的映射机制研究”[27]。

笔者在详细分析 WordNet 同义词集和 SUMO 本体概念之间映射关系的基础上，设计了一个集成算法对 WordNet 同义词集及其对应的 SUMO 本体概念进行集成，形成涵盖 WordNet 同义词集与 SUMO 本体概念一一映射关系的同义词集－概念两层索引库。集成算法的基本流程及核心代码参见笔者在《现代图书情报技术》上发表的专题文章“基于 SUMO 和 WordNet 本体集成的文本分类模型研究”[28]。集成算法输出的同义词集－概念两层索引库表的部分结果如图 2 所示。

synset	concept
a:1:{i:0;S:15:"physical";}	Physical
a:2:{i:0;S:11:"abstraction";i:1;S:15:"abstract_ent...	Abstract
a:1:{i:0;S:5:"thing";}	CorpuscularObject
a:2:{i:0;S:6:"object";i:1;S:15:"physical_object";}	CorpuscularObject
a:2:{i:0;S:5:"whole";i:1;S:4:"unit";}	CorpuscularObject
a:1:{i:0;S:8:"congener";}	familyRelation
a:2:{i:0;S:12:"living_thing";i:1;S:13:"animate_thi...	CorpuscularObject
a:2:{i:0;S:8:"organism";i:1;S:5:"being";}	Organism
a:1:{i:0;S:7:"benthos";}	Organism

图2 同义词集－概念两层索引库表的部分结果

2.1.2 同义词集－概念－本体片段三层索引库构建

同义词集－概念－本体片段三层索引库构建主要在同义词集－概念两层索引库的基础上，引入大型本体文档切割技术构建的，目的是为了向用户展示直观、清晰、大小范围适中的本体概念语义图，以防止显示的本体概念语义图过大、涉及的概念过多而造成前台信息检索界面显示不全，用户查找相关概念困难的问题。本体切割技术就是将一些体积过于庞大，难以有效展示和利用的大型本体文档切割为大小适度、易于浏览使用的小型本体片段，主要是为了解决目前大型本体利用方面存在的几个问题：（1）难以有效维护和验证。本体作为一个领域的共享概念体系，需要经得起不同领域专家的验证，以保证其一致性，然而目前一些大型本体库中概念及其语义关系非常多，如 SUMO 本体拥有 20,000 个概念和700,000个公理，如此庞大数量的概念和公理，单凭一个人是无法有效维护和验证的，需要众多领域专家学者的参与。（2）难以全面展示和有效利用。一个大型本体文档包含的概念及其语义关系过多，在目前电脑屏幕大小的限制下，很难利用可视化程序将其全方位地展示给用户，即使能够全方位地展示给用户，也会让用户眼花缭乱，并且推理起来非常困难。通常不同领域、不同主题需求的用户对本体的利用程度是不同的，比如计算机领域的用户可能仅仅需要用到包含计算机相关概念及

其语义关系的那部分本体;体育领域的用户可能仅仅需要体育方面的本体,而体育领域中的一个乒乓球运动员可能仅仅需要用到有关乒乓球的一个小小的本体片段即可。如图3所展示的从大型本体概念语义图中切割出的一个非常小的车辆本体片段的概念语义关系图,从此图中,用户很容易就能找到车辆及相关概念主题的内容。

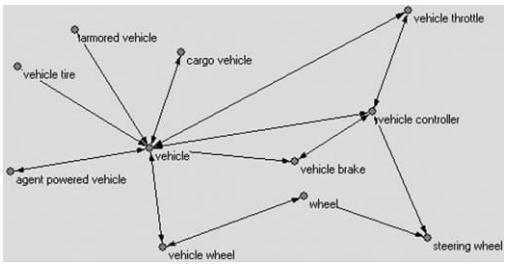

图3 车辆本体片段的概念语义图

目前国内对本体切割技术的研究基本上是空白的,国外在这方面研究比较深入的是荷兰阿姆斯特丹自由大学的两位学者Stuckenschmidt H和Klein M,他们提出基于本体类层次结构的大型本体分割方法[29],该方法首先利用SWI-Prolog Semantic Web Library[30]读入OWL或RDF schema格式的本体文件,并利用社会网络分析工具Pajek[31]生成本体的类层次结构图即本体概念语义图,同时计算类或概念结点之间的依赖程度即关系密切程度,然后基于具有依赖关系的本体概念语义图,设置本体片段大小(Maximum island size)即本体片段包含的最大概念数目,最后由孤岛算法(island algorithm)根据概念之间语义关系的密切程度,将大型本体概念语义图切割成一系列概念语义关系密切、主题相近且在人类视觉接受范围内的小规模本体概念语义图。

笔者主要采用Stuckenschmidt H和Klein M两位学者设计的大型本体分割方法,利用他们开发的本体分割工具Pato[32]对SUMO本体概念语义图进行分割,形成一系列大小适中,易于浏览使用的小规模本体片段,切割程序成功运行的结果界面如图4所示,未融合之前的类簇即本体片段的个数是519,根据片段之间相似程度或关系密切程度融合之后的本体片段个数是394。然后人工判断同义词集-概念两层索引库中概念的主题,从切割的394个小规模的本体片段中找出主题密切相关的本体片段,将其路径赋予同义词集-概念两层索引库中新增的本体片段的路径字段中,形成同义词集-概念-本体片段三层索引库。

2.2 用户查询与结果反馈模块

该模块是模型的前台用户界面,除了负责与用户进行交互,处理用户输入的查询语句,向用户反馈查询结果之外,还需与其它三个模块进行交互,因此其设计要外简内繁,即外观要简单、美观大方,方便用户浏览检索,内部程序设计要复杂一点,以便用户更快速、便捷地实现其功能,提升用户检索体验。该模块的具体实现流程:(1)调用RapidMiner中的Text组件[33]对用户输入的查询语句进行分词、停用词过滤、字符长度过滤、

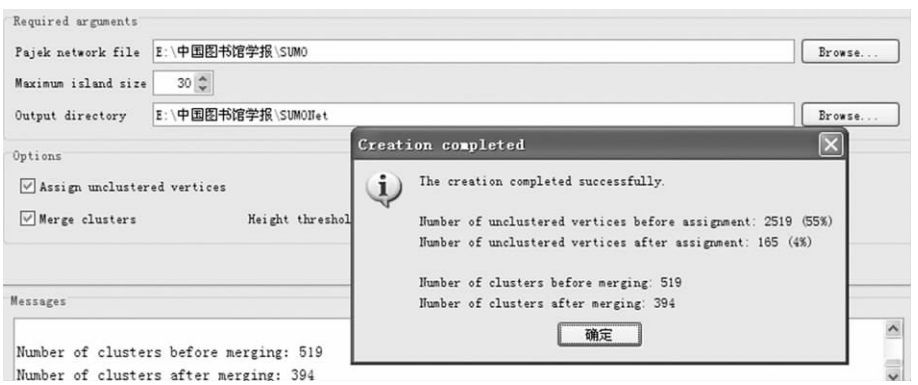

图4 切割程序成功运行的结果界面

抽取关键词，并存储于查询向量 $keywords 中。(2) 基于同义词集 - 概念 - 本体片段三层索引库对用户查询向量进行判断，首先判断查询向量 $keywords 中的每个关键词在同义词集 - 概念 - 本体片段三层索引库中的同义词集字段中是否存在。如果不存在，则直接将用户查询向量 $keywords 输送到检索引擎模块进行资源检索，并将查询向量 $keywords 中的关键词反馈给系统管理员，以便对索引库进行更新。如果存在，则判断包含每个关键词的同义词集个数 n。如果 n = 1，说明查询关键词无歧义，则直接将该同义词集追加到查询向量中，并获取同义词集对应的概念。自动调用编写的匹配程序，从该概念对应本体片段的可视化文件 * . Net 中，匹配出对其依赖程度大于或等于所设阈值的概念，即语义关系非常密切的概念，这些概念通常为该概念的同义概念、上下位概念。然后将它们加入到查询向量中，实现对查询向量的语义扩展，同时将概念对应的本体片段路径传递到融合与可视化模块中进行可视化展示，以便用户根据查询主题领域的本体概念语义图选择更专业的概念改进其检索式，进行二次检索。如果 n 大于 1，则将 n 个同义词集及其对应概念同时展示给用户，用户从中选择出能够充分表达其信息需求的同义词集及其对应概念。系统会自动将用户选择的同义词集及其对应概念加入到用户查询向量中，并自动调用匹配程序。

为了解释依赖程度这一概念术语，下面笔者展示一下车辆本体片段文件中的内容，具体如图 5 所示。其中，第 1 行表示该本体片段中的概念节点数，共 8 个；2 - 9 行表示 8 个概念节点的名称；第 10 行表示连接概念节点的弧；11 - 20 行中每行的第 3 列表示一个概念对另一个概念的依赖程度，即列 1 数字所示概念对列 2 数字所示概念的依赖程度，最后 1 列表示两个概念之间的语义关系。从图 5 可以看出，对概念 4 “车辆 (vehicle)”依赖程度最高即语义关联最密切的概念是“货车 (cargo vehicle)”。

```
1  *Vertices 8
2  1 "agent powered vehicle" ellipse
3  2 "armored vehicle" ellipse
4  3 "cargo vehicle" ellipse
5  4 "vehicle" ellipse
6  5 "vehicle brake" ellipse
7  6 "vehicle controller" ellipse
8  7 "vehicle throttle" ellipse
9  8 "vehicle tire" ellipse
10 *Arcs
11 1 4 0.2500000 1 "isa"
12 4 1 0.0416667 1 "isa"
13 2 4 0.3333333 1 "isa"
14 4 2 0.0416667 1 "isa"
15 3 4 0.6666667 1 "isa"
16 4 3 0.0416667 1 "isa"
17 5 6 0.5000000 1 "isa"
18 6 5 0.1666667 1 "isa"
19 7 6 0.5000000 1 "isa"
20 6 7 0.1666667 1 "isa"
```

图 5 车辆本体片段文件中的内容

2.3 融合与可视化模块

该模块主要负责对用户检索主题相关概念的本体片段进行融合和可视化展示，以供用户可视化检索和语义扩展检索。该模块根据用户查询与结果反馈模块传递过来的本体片段路径参数的个数进行判断：如果路径参数个数等于 1，则直接调用本体分割工具 Pato 源程序 Graph. java 文件中的 Graph. class 类将路径对应的本体片段可视化成本体概念语义图，并利用 Ajax 技术无刷新地将概念语义图传递到前台用户检索界面；如果路径参数个数大于 1，则调用 Pato 源程序 Merger. java 文件中的 Merger. class 类对相应的本体片段文件进行融合，合并成一个本体文件进行可视化展示。用户根据传递过来的相关本体概念语义图选择与检索领域相关度更高、更专业的概念进行语义扩展检索。

2.4 检索引擎模块

由于在用户查询和结果反馈模块中已经对用户查询向量进行判断和语义扩展，因此该模块除了负责接收从用户查询和结果反馈模块中传递过来的用户查询向量进行资源检索外，主要负责实现基于用户在概念语义图中的操作对用户查询向量进行改进和二次检索的功能。该功能实现的具体流程：首先根据用户在本体概念语义图中的操作，自动调用匹配程序从本体概念语义图 * . Net 文件中获取与用户点击概念语义关联度大于或等于一定阈值的概念，同时从同义词集 - 概念 - 本体片段三层索引库中匹配出用户所点击概念对应的同义词集，然后将上述概念和同义

词集加入到用户原始查询表达式中，实现用户查询的二次语义扩展查询。

3 系统实验

基于上述各模块的实现流程和方法，笔者基于 PHP + MySQL 架构开发出一个简单的原型系统对模型的性能进行实验评估。系统原型界面如图 6 所示，系统中显示的是昆虫“insect”在同义词集 - 概念 - 本体片段索引库中的索引结果，从索引结果数量可以看出，昆虫的同义词集项多达 31 项，包括各种类型的昆虫和昆虫的各种属性特征，用户可以通过此索引结果很快速地了解该领域的专业词汇，并选择出能够精确表达自己需求的表达式。通常用户在检索时遵循最省力法则，总是想输入最少的单词检索到最有用的资源，然而事实并非如此，因此同义词集 - 概念索引结果的加入，可以增强用户检索体验和检索效果。

图 6 系统原型界面

3.1 实验设计

实验采用的本体集： WordNet 同义词典、SUMO 本体及两者之间的映射文件。

实验条件：（台式机）处理器为英特尔® 酷睿™ i3 2105 3.1GHz，内存为 4G DDR3，概念选择的阈值设置为 0.5。

实验检索表达式及评估方式： 笔者从 Yahoo 问题库（英文版）[34] 中抽取 5 个典型的、用户难以回答的、与 SUMO 本体概念涵盖领域相关的问题作为检索实例（见表 1），调用 Google 检索入口进行原始检索和语义扩展检索实验，并基于检索结果，利用图表对两种检索方式在 Google 检索平台中的检准率和平均检准率进行对比分析。为准确考察模型性能，我们首先请 10 位专业人士（情报学硕士、博士和科技查新人员）对检索结果中与检索实例主题密切相关的结果进行判定，以便更精确地确定检索结果的检准率指标。

实验数据集： 基于表 1 检索实例主题设计原始检索表达式，基于三层索引库扩展的语义扩展检索表达式，利用 Google 搜索引擎进行检索，将检索结果作为实验数据集。由于检索结果过多，并且无法获取 Google 搜索引擎后台数据中数据集的详细情况，因此，我们仅采用检索结果中的前 100 条记录去评估模型的性能。

评估指标： 采用检准率和平均检准率两个指标评估模型的性能。检准率记为 Pre@100，表示检索结果中前一百条检索记录的检准率。计算公式：$Pre@100 = N_r/100$，其中 N_r 表示前一百条检索记录中与用户检索主题相关的记录个数。平均检准率记为 $AvgPre@100$，表示五组检索表达式的平均检准率。计算公式：$AvgPre@100 = \sum_{i=1}^{5} Pre@100_i/5$，

其中 $\sum_{i=1}^{5} \text{Pre@100}_i$ 表示五组检索表达式的检准率之和。

表1 检索实例及其含义

查询 ID	检索实例	实例含义
1	What type or size of horse for large framed beginer rider?	体积比较大且刚学骑马的人需要什么类型和大小的马匹?
2	What is the best vehicle to get for a 26 year old stay at home with 2 daughters?	对于一个26岁的居家主妇和2个女儿的三口之家,最适合她们的车是什么?
3	What is the latest method of waste disposal without harming the environment?	不会危害环境的最新废水处理方法是什么?
4	How do we get rid of cookies,spyware and adware from a website that put them in my computers memory?	我们如何摆脱我们正浏览的网站放入我们电脑内存中的cookies、间谍软件和广告软件?
5	How do I go about introducing a young mouse to my older mouse?	如何将一个年轻的老鼠引见给一个年长的老鼠?

3.2 实验结果及分析

实验结果见表2。表2表示五组原始查询表达式和扩展查询表达式在Google搜索引擎中前100条检索记录中的检准率和平均检准率。从表2可以看出,用户原始查询表达式在使用了我们提出的语义扩展方法进行语义扩展后,检准率Pre@100和平均检准率AvgPre@100都有一定程度的提高。比如,平均检准率AvgPre@100在经过语义扩展后,比原来提高了18.20%。在检准率指标中,检索实例2的检准率最低,它原始检索表达式是"vehicle suit three people",检准率仅为3%,并且都是弱相关的记录。经过语义扩展程序扩展后的检索表达式是"type OR kind OR size OR brand OR price car OR vehicle OR automobile OR motorcar suit OR fit OR meet 'three people'",检准率提高到28%,不过尽管如此,检索结果并不令人满意,与用户检索主题密切相关的记录较少。从检索实例2的语义扩展检索表达式可以看出,原始检索关键词的同义词集和关键词对应概念的属性都是用"OR"连接。另外,对于特别抽象的概念,笔者尽量通过增大概念选择阈值,避免其加入到检索式中。这样做不仅可以避免以牺牲检全率为代价换取检准率的提高,还可以通过检索主题相关词汇的密集出现,提高检索结果的命中率,增强检索结果与用户检索主题的相关程度。

当然,由于集成的同义词集-概念-本体片段三层索引库只是一个雏形,有些同义词集涵盖的同义词并不全面和集中,在检索中有些同义词在索引库中并不存在。为说明模型的性能,笔者将索引库中需要用到但不全的同义词集进行了补充和更新。不过如果将该索引库更好地应用于网络信息检索中,则需要更多用户、学者、专家对该索引库进行补充和更新。

表2 原始查询和扩展查询分别在Google实验平台中的检准率和平均检准率

查询 ID	原始查询表达式	扩展查询表达式
	Pre@100	Pre@100
1	70%	88%
2	3%	28%
3	76%	87%
4	89%	97%
5	12%	41%
AvgPre@100	50.00%	68.20%

4 结语

文章针对传统信息检索方式存在的问题和目前语义检索中存在的不足,提出基于WordNet和SUMO本体集成的自动语义检索及可视化模型。模型重点研究了如下内容:(1)利用WordNet同义词集和SUMO本体概念之间的语义映射关系,编写正则表达式,对两大本体库的同义词集部分和概念部分进行集成,构建同义词集-概念两层索引库,并成功应用于所设计的自动语义检索及可视化模型中。集成方法简单有效,从而推动本体集成领域研究从单纯的集成理论、模型和

框架研究深入到本体集成在具体领域的应用研究。(2) 本体集成致力于通过集成使本体的涵盖领域更广,概念及其语义关系更丰富,或者某个领域需要用到不同本体库的不同部分时,将它们有效融合起来。本体分割理论致力于将一个大型本体分割成一系列涵盖领域非常小,概念及其语义关系非常少的小本体,以便更有效地维护、展示和利用它们。本文利用本体集成技术构建同义词集 - 概念两层索引库,在此基础上利用本体分割技术对概念所在的大型本体库进行分割,形成同义词集 - 概念 - 本体片段三层索引库。将两个看似相互矛盾的技术融合起来,可以起到推动两个方向融合与发展的目的。(3) 利用同义词集 - 概念 - 本体片段三层索引库输出与用户检索关键词相关的所有主题领域的同义词集及其对应概念索引,可以让用户更精确地选择检索的主题领域,并利用用户选择的同义词集及其对应概念、与此概念语义关联比较密切的同义概念和上下位概念对用户查询关键词进行语义扩展。避免了用户盲目输入检索词,检索词过少、不专业、无法准确描述用户检索需求,以及检索词的同义、多义等现象造成检索结果多,但相关的信息过少甚至没有的情形,能够有效降低获取用户查询意图和语义扩展的难度和复杂性,提高语义检索系统的性能和自动化水平。同时通过可视化技术输出与用户检索关键词相关的本体概念语义图,实现资源的可视化和个性化检索,可以有效解决目前语义检索领域在利用本体时,无法将用户检索主题领域的本体概念语义关系清晰地、全方位地展示给用户的问题,提升用户检索体验。

不足之处: 系统原型只是一个实验系统,功能单一,仅仅通过调用 Google 检索入口进行实验评估,无法获取数据的全貌,因此仅仅能评估模型的准确率; WordNet 同义词集和 SUMO 本体更新速度慢,不能随着网络动态更新,很多新出现的词汇和用户查询时在同义词集 - 概念 - 本体片段三层索引库中不存在的词汇,都无法及时纳入索引库中; 索引库是英文的,因此模型目前只能应用于英文网络信息检索领域。

未来研究方向: 对各模块功能进行无缝集成,引入海量信息采集、分类与索引模块,构成一个功能强大、能够应用于实践的语义分类与检索系统; 编写一个动态更新程序,对索引库进行动态更新; 集成中文领域的本体如 HowNet 等,构建面向中文领域的索引库,对中文自动语义检索及可视化模型进行研究和实现。

参考文献

1 Suggested Upper Merged Ontology (SUMO) [EB/OL]. [2010 - 05 - 10]. http://www.ontologyportal.org/.

2 About WordNet [EB/OL]. [2010 - 05 - 10]. http://wordnet.princeton.edu/.

3 DBpedia [EB/OL]. [2010 - 05 - 10]. http://blog.dbpedia.org/.

4 OpenCyc Documentation [EB/OL]. [2010 - 05 - 11]. http://www.opencyc.org/doc.

5 董振东,董强. 关于知网—中文信息结构库 [EB/OL]. [2010 - 05 - 11]. http://www.keenage.com/html/e_index.html.

6 The Enterprise Ontology [EB/OL]. [2010 - 05 - 12]. http://www.aiai.ed.ac.uk/project/enterprise/ontology.html.

7 王效岳,胡泽文,白如江,李玉平. 本体集成: 概念、过程、工具与方法综述 [J]. 图书情报工作,2011,55 (16): 119 - 125.

8 Prompt [EB/OL]. [2010 - 06 - 01]. http://protege.stanford.edu/plugins/prompt/prompt.html.

9 OntoMerge [EB/OL]. [2010 - 06 - 01]. http://cs-www.cs.yale.edu/homes/dvm/daml/ontology-translation.html.

10 StaabS, Studer R. Handbook on Ontologies in Information Systems [M]. Germany: Springer-Verlag, 2004: 397 - 416.

11 Kiryakov A, Kiril l. S, Dimitrov M. OntoMap: portal for upper-level ontologies [C]. Proceedings of Formal Ontology in Information Systems: Collected Papers from the Second Inter-

national Conference, Ogunquit, ME, United states,2001:47 - 58.

12 COMA + + [EB/OL]. [2010 - 06 - 03]. http://dbs.uni-leipzig.de/Research/coma.html.

13 Stumme G, Maedehe A. FCA-Merge: Bottom-Up Merging of Ontologies [C]. Proceedings of the Seventeenth International Conference on Artificial Intelligence (IJCA101), Seattle, WA, USA, 2001:225 - 230.

14 Hitzler P, krotzsch M, Ehrig M, et al. what Is Ontology Merging-A Category-Theoretical Perspective using Pushouts [J]. American Association for Artificial Intelligence,2005,6 (6): 104 - 107.

15 卢胜军,李法勇,钱建军等. WCONS +:一种基于 WCONS 的本体集成方法 [J]. 现代图书情报技术,2008 (2):18 - 22.

16 张忠平,赵海亮,张志惠. 基于 OWL 的本体集成 [J]. 计算机应用,2008 (28):10 - 14.

17 张忠平,赵海亮,田淑霞. 基于 RDFS 的本体集成方法 [J]. 计算机工程与应用,2008 (15):131 - 141.

18 HuW, Qu Y. Z. Falcon-AO: A practical ontology matching system [J]. Journal of Web Semantics,2008,6 (03):237 - 239.

19 魏哲雄,冯志勇. 基于字典技术的本体整合系统 [J]. 计算机应用,2007 (2):428—430.

20 Park L. A. F,Ramamohanarao K. Efficient storage and retrieval of probabilistic latent semantic information for information retrieval [J]. VLDB Journal,2009,18 (1): 141 - 155.

21 MehulB,Wenny R,Prakash S. S ,Carlo W. Ontology driven semantic profiling and retrieval in medical information systems [J]. Journal of Web Semantics,2009,7 (4):317 - 331.

22 周文,龚礼明,蒋岚. 隐含语义检索及中文样本分析实例 [J]. 计算机应用,2004 (4):273 - 276

23 何琳,侯汉清,杜慧平. 一种基于领域本体的语义检索系统的设计与实现 [J]. 图书情报工作,2008 (8):85 - 88.

24 Adam P,Ian N,and John L. The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications [C]. In: Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web, Edmonton, Canada,2002:2002.

25 George A M. WordNet: A Lexical Database for English [J]. Communications of the ACM, 1995,38 (11): 39 - 41.

26 Suggested Upper Merged Ontology (SUMO) [EB/OL]. [2010 - 06 - 10]. http://www.ontologyportal.org/.

27 王效岳,胡泽文,白如江. WordNet 与 SUMO 本体之间的映射机制研究 [J]. 现代图书情报技术,2011 (1):22 - 30.

28 胡泽文,王效岳,白如江. 基于 SUMO 和 WordNet 本体集成的文本分类模型研究 [J]. 现代图书情报技术,2011 (01): 31 - 38.

29 StuckenschmidtH, Klein M. Structure-Based Partitioning of Large Concept Hierarchies [J]. Lecture Notes in Computer Science, 2004 (3298):289 - 303.

30 Wielemaker J. ,Schreiber G. ,Wielinga B. : Prolog-based infrastructure for RDF: performance and scalability [M]. In Fensel, D. , Sycara, K. , Mylopoulos, J. , eds. : The Semantic Web-Proceedings ISWC'03, Sanibel Island, Florida, Berlin, Germany, Springer Verlag (2003) 644 - 658 LNCS 2870.

31 Batagelj V. ,Mrvar A. : Pajek-analysis and visualization of large networks [M]. In Jnger M. , Mutzel P. , eds. : Graph Drawing Software. Springer,2003: 77 - 103.

32 Structure-based Ontology Partitioning [EB/OL]. [2010 - 06 - 10]. http://swserver.cs.vu.nl/partitioning/.

33 Download RapidMiner Extensions [EB/OL]. [2010-06-14]. http://rapid-(下转第 91 页)

10 吴建中. 21 世纪图书馆新论[M]. 上海: 上海科学技术文献出版社,1998.
11 王子舟. 图书馆学基础教程[M]. 武汉: 武汉大学出版社,2003.
12 阮冈纳赞. 图书馆学五定律[M]. 夏云,译. 北京: 书目文献出版社,1988.
13 宓浩. 图书馆学原理[M]. 上海: 华东师范大学出版社,1988.
14 袁咏秋. 外国图书馆学名著选读[M]. 北京: 北京大学出版社,1988.
15 孟广均. 国外图书馆学情报学研究进展[M]. 北京: 北京图书馆出版社,1999.
16 中国图书馆分类法编辑委员会. 中国图书馆分类法(第4版)[M]. 北京: 北京图书馆出版社,1999.
17 初景利. 复合图书馆的概念及发展构想[J]. 中国图书馆学报,2001(3):3-6.
18 吴建中. DC 元数据[M]. 上海: 上海科学技术文献出版社,2000.
19 高文. 数字图书馆——原理与技术实现[M]. 北京: 清华大学出版社,2000.
20 刘炜. 数字图书馆引论[M]. 上海: 上海科学技术文献出版社,2001.
21 Arms WY. 数字图书馆概论[M]. 施伯乐,译. 北京: 电子工业出版社,2001.
22 吴志荣. 数字图书馆——从理念走向现实[M]. 上海: 学林出版社,2000.
23 张晓林. 数字图书馆机制的范式演变及其挑战[J]. 中国图书馆学报,2001(6):3-8,17.
24 蒋永福. 维护知识自由: 图书馆职业的核心价值[J]. 2003(6):1-4.
25 黄宗忠. 论图书馆核心价值(上)[J]. 图书馆论坛,2007(6):3-8.
26 黄宗忠. 论图书馆核心价值(下)[J]. 图书馆论坛,2008(1):1-3.
27 范并思,胡小菁. 图书馆2.0: 构建新的图书馆服务[J]. 大学图书馆学报,2006(1):2-7.

(宗乾进 南京大学信息管理学院2010级博士研究生,袁勤俭 教授 南京大学信息管理学院,沈洪洲 南京大学信息管理学院2010级博士研究生)

收稿日期: 2011-07-19

(上接第32页)

i.com/content/view/55/85/.
34 YAHOO! ANSWERS[EB/OL]. [2010-06-19]. http://global.ard.yahoo.com/SIG=15oivh6vm/M=650008.12773057.13811805.8356343/D=know/S=396545059:HEAD/Y=YAHOO/EXP=1306211143/L=G5MTw0S00ia336UITdsWdADhym7Ro03bFyYAAYik/B=3dG4AGKJiTU-/J=1306203943095169/K=8B7ZdM_qNTG8iqBcruKEPA/A=5856910/R=11/SIG=10qrh7h9e/* http://answers.yahoo.com.

(胡泽文 南京大学信息管理学院2011级博士研究生)

收稿日期: 2011-10-01