《智能信息处理》课程考试

基于本体增强语义信息检索

李安宁

考核	到课[10]	作业[20]	考试[70]	课程成绩[100]
得分				

2021 年 12 月 08 日

基于本体增强语义信息检索

李安宁

（大连海事大学 信息科学技术学院,辽宁 大连 116026）

摘 要 信息检索(Information Retrieval)是从非结构化文档中识别和检索存储在其中的特定信息的过程。传统基于关键字的信息检索不能将原始数据转换为包含语义的表示数据，提供了有限的与用户需求相关联的信息。为了解决其局限性，语义网(SemanticWeb，SW)和本体论(Ontology)概念应运而生，从而实现了基于语义而不是字符串进行信息搜索，本体的引入有利的推动了构建新一代的语义 Web，提高了机器对语义 Web 中概念理解。本文描述了基于本体的语义信息检索模型，通过在本体中定义术语并将其作为元数据，以一种有力形式标记 web 的内容来构建用于信息检索的语义索引项，有效的提高了信息检索的效果。

关键词 信息检索；本体；数据挖掘；语义网

中图法分类号 TP311 文献标识码 A

Ontology-based enhanced semantic information retrieval

Li Anning

(school of Information Science and Technology, Dalian Maritime University, Liaoning Dalian 116026)

Abstract Information Retrieval is the process of identifying and retrieving specific information stored in an unstructured document. Traditional keyword-based information retrieval cannot convert raw data into representational data containing semantics, providing limited information associated with user needs. In order to solve its limitations, SemanticWeb (SW) and Ontology concepts emerged as the times require, so that information retrieval based on semantics rather than strings is realized. The introduction of ontology has promoted the construction of a new generation of Semantic Web. Improves machine understanding of concepts in the Semantic Web. This paper describes an ontology-based semantic information retrieval model. By defining terms in ontology and using them as metadata, the semantic index items for information retrieval are constructed by marking the content of the web in a powerful form, which effectively improves information retrieval effect.

Keywords Information Retrieval (IR); Ontology;Data Mining; Semantic Web (SW)

1 引 言

随着信息化时代的到来，信息检索[1]技术的研究不断深入推进，以便为下游任务提供更加及时、精准、有效的数据和信息。信息检索是指根据用户需求从海量的、非结构化文档中发现和提取相关信息。用户仅仅通过键入与需求相关的关键字，互联网就能通过链接到各种数据和文档超链接获取信息，即时送入用户界面。但 web 的信息检索结果却往往不令人满意，检索到的信息可能是相关的，也可能是不相关的，它仅仅是根据字符串匹配来进行信息搜寻，这就导致了信息检索的结果不全面或者冗余信息过多，需要人工加以分辨和提取，大大缩减了检索效果。造成上述情况的根本原因是计算机无法理解现有 Web 网页内容的语义，计算机只是按照指令完成检索流程，但网页具体是

什么，检索到的信息与用户想要的信息关联大不大，计算机都无法理解和处理。这一弊端是传统检索技术效果不佳的重要弊端。那如何让计算机能理解网页内容，出于此目的，TimBerners-Lee 提出了语义 Web[2]。语义 Web 不是一个凭空出现、自成一体的 Web 网络，而是现有万维网的扩展。语义 Web 是一个抽象的信息集合，在该扩展中，通过在计算机内添加 Ontology 中的概念，赋予 Web 中的信息明确的含义，使得计算机从而理解文本内容，更好地与用户进行协作，完成信息检索、智能问答等任务。因此利用本体构建语义 Web 以及从语义层进行文本检索是一个更好地趋势。

2 相关概念

2.1 本体

“本体”一词来源于哲学中的术语，其本质就是一个概念的模型，表明规范化的概念，为共享的概念及其相互之间关系定义了一种形式化的、规范的、详细的说明[3]。也就是说，本体可以用于对存在的概念和关系的进行统一、正式的定义和描述。

本体在提高信息检索性能方面有着极大的优势，主要表现为以下几个方面：一是本体是元数据模式，提供了概念的模型[4]，使用本体概念可以将给定的数据转换为机器可理解的语言，使用本体可以克服了基于关键字的搜索的局限性，促进了基于语义的信息检索的出现和发展。二是利用实体可以处理事件的实例以及用户定义的概念之间的关系，并推动了语义 Web 的发展。因此。通过将语义层定义为语义实体的集合，基于语义层的背景知识就包括了实体的概念和关系，而不再是简单的词语，这样极大的促进了计算机对语义层特定实体间关系的理解、掌握领域问题的事实和规则。

2.2 信息检索

信息检索是指根据用户需求对存储在计算机中的非结构化文档进行识别和检索的过程。非结构化文档形式多种多样，主要包括视频、照片和音频等。现有的信息检索主要集中于自然语言文本的的搜索信息检索引擎的体系结构是基于本体的模型，主要由以下部分构成：①OMC(Ontology Manager 组件)：它由 Indexer、搜索引擎和 GUI 使用；②索引器:它对文档进行索引并创建元数据③搜寻引擎④GUI：支持用户的查询格式[5]。

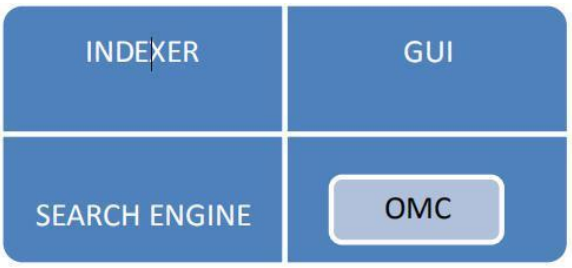


图 1 信息检索引擎体系结构

2.3 语义网

TimBernersLee 在 1996 提出了“语义网(SW)”是对现有万维网和扩展和完善，核心思想是通过机器可理解的标注语言对文档进行注释，因此实现计算机对 Web 网页内容的理解，以便更好的处理用户的需求。语义网是表达信息的框架，基于语义网该特征，可以提高信息检索的效果。SW 中用到的技术如下[6]：

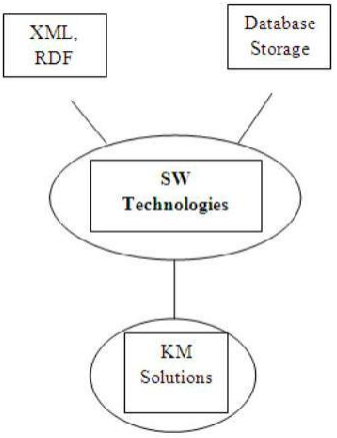


图 2 语义 Web 技术

XML：XML 是可扩展的语言，允许用户创建自己的文档标记。它为文档中的内容结构提供语法。XML Schem 它是定义 XML 文档的语言。

RDF：代表资源描述框架。它提供了一种表示数

据模型的简单的语言形式，刻画出引用对象以及他们的关系这些模型称为 RDF 模型。XML 和 RDF 都处理与其他数据有关的元数据。

3 基于本体的语义网信息检索模型

利用本体和语义 Web 推理机完成信息检索过程，需要按照以下列方式处理查询并生成结果：

首先将用户查询输入到推理引擎中，该查询经过处理后编码为可以被搜索引擎识别的文本查询，生成用于检索信息的语义标记文档[7]。

其次，经过标识后，查询将提交到一个或多个网页。这期间，一些网页的标记可能不相关或未经授权。FILTERS 用于处理这些标记并让其完全授权。过滤后，这些事实将被传递给推理引擎。

重复此过程，直到完成信息检索。

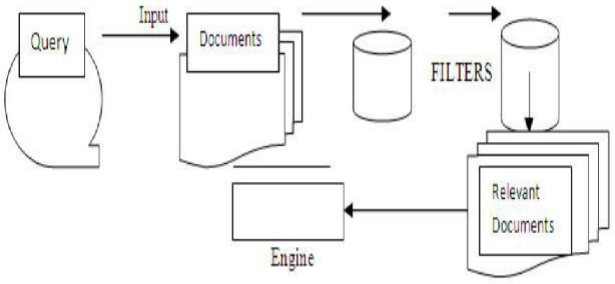


图 3 信息检索引擎体系结构

3.1 获得语义索引项

传统方法是通过文档向量之间内积的余弦来计算用户查询 q 和文档 d 之间的相似度[8]，具体公式如下:

$$Sim(q,d)=\frac{\sum_i q_i d_i}{\sqrt{\sum_i q_i^2}\ \sqrt{\sum_i d_i^2}}$$

其中，q_i 和 d_i 是两个向量表示中的权值。

根据上述公式计算出所有查询和文档间的相似度之后，将其进行排序，得到最终结果。然而，两个文档中不存在相同的词不代表两篇文档中没有关联。例如“乔布斯去世”和“苹果的价格会下降吗？”，仅仅根据相似度的计算无法得出两个文本之间的关联，因此利用本体来进行语义标注获得语义索引项才能挖掘词汇和文档间的语义相似度。

查询 q 和文档 d 在经过预处理后,抽取出索引项构成向量。因此，模型首要工作是获取索引项。将本体中定义好的术语作为元数据来标记 Web 的内容，这些语义标记就可以作为语义索引项。获得语义索引项后，中通过每个标引项的权值生成向量表示，可看作为其在文档集合中出现频率的函数 ，用 tf、idf 来计算。

3.2 生成逻辑视图

假设 x_i 是获取到的语义索引项， x_i 的等价类[9]表示为[x_i]，其中 x_{i1} 、 x_{i2} 、...∈[x_i .]，假设 x，y 和 z 表示在文档 d_j 出现的 3 个不同的字符串，将 x,y,z 分别语义标注为 x_{i1} 、 x_{i2} 、 x_{i3} 。按照该方法，虽然 x、y 和 z 是不同的字符串，但从语义的角度看，因为都用相同的语义索引项 x_i 表示，因此 x 、 y 和 z 是相等的。 假设语义网上的文档集合为 D，D={ d_1 ， d_2 ，…， d_n }(1≤j≤ n)；假设文档集合 D 的语义索引项的所有等价类集合：{[x_1]，[x_2]，…[x_t]}，在文档 dj 的所有语义标记中，来自等价类[x_i](1≤i≤t)的所有元素的总频率为 f_{ij} 。(即，从等价类[x_i]中的所有元素出现在文档 d_j 的所有语义标记中的总次数[x_i])，如表 1 所示。

Equivalent class	d_1	d_2	…	d_i	…	d_n
$[X_1]$	f_{11}	f_{12}	…	f_{1i}	…	f_{1n}
$[X_2]$	f_{21}	f_{22}	…	f_{2i}	…	f_{2n}
…	…	…	…	…	…	…
$[X_t]$	f_{t1}	f_{t2}	…	f_{ti}	…	f_{tn}
…	…	…	…	…	…	…
$[X_t]$	f_{t1}	f_{t2}	…	f_{ti}	…	f_{tn}

表 1 等价类中的所有元素出现在文档的所有语义标记中的总频率

根据以下公式可以计算出所有的权重（n_i 是包含等价类[xi]的元素的文档数）：

$$\begin{cases} tf_{ij} = \dfrac{f_{ij}}{\max\{f_{1j}, f_{2j}, \ldots, f_{ij}\}} \\ idf_i = \log_2(\dfrac{n}{n_i}) \\ \omega_{ij} = tf_{ij} \times idf_i \end{cases}$$

由权重可以得到 D 中每个文档的逻辑视图：$d_j \cong (\omega_{1j}, \omega_{2j}, \ldots\ldots\omega_{ij})$，通过把查询中的索引项映射到本体，这样模型就可以得到索引项的在本体中语义，从而可以很好地表示文档。

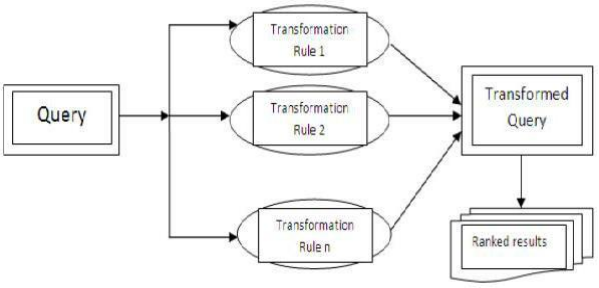
对于用户查询信息 q，利用语义索引项的等价类（{[X1]，[X2]，…）获得 q 的逻辑视图 $q \cong (\omega_{qj}, \omega_{qj}, \ldots\ldots\omega_{qj})$，通过该逻辑视图，用户需求信息 q 也注入了语义。

3.3 排序函数

利用排序函数计算用户查询 q 和文档 d_j 之间的相似度。即排序函数显示出针对某一特定的查询 q，文档 d_j 排名情况。排序函数为：

$$\frac{1}{\sqrt{\sum_{i=1}^{t}(\omega_{iq} - \omega_{ij})^2}}$$

以下是基于本体的信息检索模型的各个关键部分及处理流程：



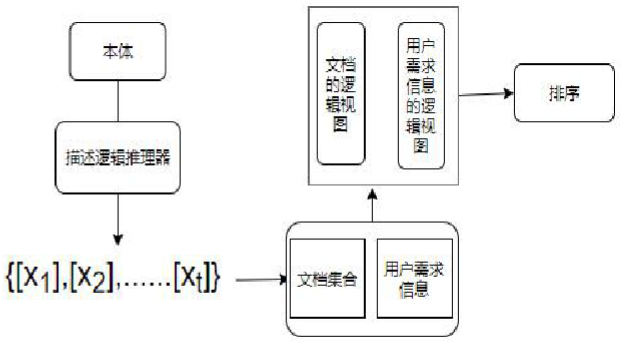


图 4 基于本体的信息检索模型

4 模型评价

基于本体增强语义信息检索模型主要有以下几点优势：

首先模型可以利用本体中的概念生成的语义索引项根据词频 tf、逆向文件频率 idf 为查询和文档中的索引项初始化权值，并根据标引项之间的语义关系进行权值调整，考虑了概念之间词义的相似度，从而可提高查询精度。

另外，该模型还可以利用排序算法对生成的用户查询信息和文档信息的逻辑视图的相似度的大小进行排序，从而从语义层面考虑相似度，远比通过文档向量之间内积的余弦来计算用户查询和文档之间的相似度要可靠的多。

5 结 语

现如今，问答系统、智能决策等各个应用领域对信息检索的速率和质量提出了更高的要求。传统基于关键字的信息检索模型仅仅是将字符串进行匹配[10]，而无法从语义层面对查询信息和网页信息进行理解和处理，因而检索效果不佳，本文提出了基于本体的语义检索模型，将本体中定义好的术语作为元数据来标记 Web 的内容，这些语义标记就可以作为语义索引项。由等价类和语义索引项计算所有元素出现在文档的所有语义标记中的总频率及权重，生成已经注入语义的用户查询信息和文档的逻辑视

图，最后，利用排序函数计算用户查询和文档之间的相似度。基于此模型，信息检索实现了对信息进行语义层面的处理，使得计算机能够理解用户查询和网页内容，更好地与用户交互，极大地提升了检索效果。

参考文献

[1]Aline Chevalier,Aurélie Dommes,Jean-Claude Marquié. Strategy and accuracy during information search on the Web: Effects of age and complexity of the search questions[J]. Computers in Human Behavior,2015,5

[2]刘柏嵩. 基于知识的语义网:概念、技术及挑战[J]. 中国图书馆学报, 2003(02):17-20.

[3]Peng Jiajie,Wang Honggang,Lu Junya,Hui Weiwei,Wang Yadong,Shang Xuequn. Identifying term relations cross different gene ontology categories.[J]. BMC bioinformatics,2017,18(Suppl 16).

[4]刘宇松.本体构建方法和开发工具研究[J].现代情报, 2009,29(09):17-24.

[5]王元卓,贾岩涛,刘大伟,靳小龙,程学旗.基于开放网络知识的信息检索与数据挖掘[J].计算机研究与发展,2015,52(02):456-474.

[6]Jafarpour Borna,Abidi Samina Raza,Abidi Syed Sibte Raza. Exploiting Semantic Web Technologies to Develop OWL-Based Clinical Practice Guideline Execution Engines.[J]. IEEE journal of biomedical and health informatics,2016,20(1).

[7]曲佳彬.基于本体的语义信息检索分析[J].技术与市场,2010,17(12):84-85.

[8]李晓红.基于本体技术的语义检索及其语义相似度分析[J].电子技术与软件工程,2017(01):187.

[9]唐守利. 基于本体的云服务语义检索模型研究[D].吉林大学,2016.

[10]阮春阳. 基于本体的语义相似度研究[D]. 郑州大学, 2016.