计算机研究与发展 ISSN 1000-1239/CN 11-1777/TP
Journal of Computer Research and Development 47(2): 255-263, 2010

基于 RDF 句子的语义网文档搜索

吴鸿汉 瞿裕忠 李慧颖
(东南大学计算机科学与工程学院 南京 210096)
(hhwu@seu.edu.cn)

Searching Semantic Web Documents Based on RDF Sentences

Wu Honghan, Qu Yuzhong, and Li Huiying
(*School of Computer Science and Engineering, Southeast University, Nanjing* 210096)

Abstract Keyword-based semantic Web document search is one of the most efficient approaches to find semantic Web data. Most existing approaches are based on traditional IR technologies, in which documents are modeled as bag of words. The authors identify the difficulties of these technologies in processing RDF documents, namely, preserving data structures, processing linked data and generating snippets. An approach is proposed to model the semantic Web document from its abstract syntax: RDF graph. In this approach, a document is modeled as a set of RDF sentences. It preserves the RDF sentence-based structures in the processes of document analyzing and indexing. The authoritative descriptions of named resources are also introduced and it enables the linked data across document boundaries to be searchable. Furthermore, to help users quickly determine whether one result is relevant or not, The traditional inverse index structure is extended to enable more understandable snippet extraction from matched documents. Experiments on real world data show that this approach can significantly improve the precision and recall of semantic Web document search. The precision at top one result is improved up to 19% and a steady improvement (near 10%) is observed. According to 50 random queries, the recall increases up to 60% averagely. Remarkable improvements in system usability are also obtained.

Key words semantic Web; search engine; RDF document search; RDF sentence; snippet generation

摘 要 语义网文档搜索是发现语义网数据的重要手段.针对传统信息检索方法的不足,提出基于 RDF 句子的文档词向量构建方法.首先,文档被看作 RDF 句子的集合,从而在文档分析和索引时能够保留基于 RDF 句子的结构信息.其次,引入资源的权威描述的定义,能够跨越文档边界搜索到语义网中互连的数据.此外,扩展了传统的倒排索引结构,使得系统能够提取出更加便于阅读和理解的片段.在大规模真实数据集上的实验表明,该方法可以显著地提高文档检索的效率,在可用性上具有明显的提升.

关键词 语义网;搜索引擎;语义网文档搜索;RDF 句子;片段提取
中图法分类号 TP393.09

在语义网中,通过文档的方式发布 RDF 数据是语义网数据发布的重要手段之一.这些语义网文档是语义网数据重用和共享的重要信息来源.语义网文档搜索服务是访问这些信息源的最快捷和最方便

收稿日期:2008-11-05;修回日期:2009-10-22
基金项目:国家自然科学基金项目(60773106);江苏省自然科学基金项目(BK2008290)

的手段之一. 因此, 很多语义网搜索引擎都提供了语义网文档的搜索服务, 包括 Swoogle[1], Sindice[2] 和 Watson[3] 等.

目前大部分的语义网文档搜索服务都直接采用传统信息检索的技术对语义网文档进行分析和编排索引. 比如 Swoogle 系统采用基于 n-gram 的技术对语义网文档进行分析和处理[4]. 然而, 语义网文档与传统的 Web 文档在语言模型以及针对的应用上都存在着很大的差异. 传统的 Web 文档采用 HTML/XHTML 语言, 对自然语言的显示进行标记, 供用户进行阅读和浏览. 而 RDF 是基于三元组的结构化的数据模型, 目标是实现 Web 范围内信息的集成和共享. 这种差别使得直接采用传统信息检索技术不能很好地满足语义网文档搜索的需求. 其主要的不足体现在以下 3 个方面.

首先, 传统的信息检索技术在处理语义网文档时会丢失语义网文档所表达的结构信息. 在传统的信息检索模型中, 词频是相关度考量的重要因素之一. 然而, 这种方法对于语义网文档并不(直接)适用. 图 1 显示了一个语义网文档的片段. 其表达的含义是 Tim Berners-Lee 认识 ruthdhan 和 Tim Berners-Lee 认识 crowell. 从语义来看, 该段数据更多的描述的是 Tim 的信息. 然而, 单纯从词汇袋的模型来看, Berners-Lee 的词频为 1, ruthdhan 的词频为 2. 因此, 传统信息检索的方法不能够处理 RDF 数据的结构信息, 进而难以很好地反映数据的含义.

```
—con: Male rdf: about = “ http:// www. w3. org/ People/ Berners-Lee/
  card# i”—
  —knows rdf: resource= “ http:// web. mit. edu/ ruthdhan/ www/ foaf.
    rdf# ruthdhan”/—
  —knows rdf: resource= “ http:// people. csail. mit. edu/ crowell/ foaf.
    rdf# crowell”/—
—/con: Male—
```

Fig. 1 A snatch of Tim Berners-Lee's FOAF document.

图 1 Tim Berners-Lee 的 FOAF 文档片段

其次, 传统信息检索方法搜索不到语义网中通过 URI 重用带来的互联数据. 语义网中实体的指称是基于 URI 的, 而 URI 本身并不一定携带其所指称的实体的描述. 在图 2 的例子中, Tim Berners-Lee 在其 FOAF 文档中定义了自己的 URI; Lalana Kagal 在其文档中重用了这个 URI. 对于直接采用传统的信息检索技术的系统而言, 第 2 个文档不会包含 Tim Berners-Lee 的名字信息: Tim. 因此, 采用传统信息检索技术不能处理这种跨越文档的 URI 重用的情况, 从而降低了系统的召回率.

```
□Tim Berners-Lee's FOAF Doc (http:// www. w3. org/ People/
  Berners-Lee/ card. rdf) - Define URI:
—con: Male rdf: about = “ http:// www. w3. org/ People/ Berners-Lee/
  card# i”—
  —s: label—Tim Berners-Lee—/s: label—
—/con—
□Lalana Kagal's FOAF Doc (http:// people. csail. mit. edu/ lkagal/
  foaf) - Reuse URI:
—foaf: Person rdf: ID= “ me”—
  —foaf: knows rdf: resource= “ http:// www. w3. org/ People/ Berners-
    Lee/ card# i”/—
—/foaf: Person—
```

Fig. 2 An example of URI reuse across documents.

图 2 跨文档的 URI 重用示例

最后, 传统方法很难从语义网文档中提取适合理解的、语义完整的内容片段来帮助用户判断结果文档的相关性. 结果文档的片段提取(snippet generation)能够帮助用户快速定位到需要的文档. 对于传统文档而言, 通常的做法是以查询关键词出现的位置为锚点, 提取出锚点前后一定长度的字串. 而对于语义网文档而言, 这种做法很难提取到可以让用户理解的片段. 一方面是因为 RDF 模型中采用 URI 来指称资源, 这种指代会带来局部信息缺失. 另一方面是由于语义网文档是其所描述的 RDF Graph 的一个序列化表示, 这个序列化的过程可能把 RDF Graph 中语义相近的部分物理上分开存放. 出于上述原因, 目前所知的语义网文档的搜索引擎都没有提供适合阅读的、语义完整的关键词匹配的证据信息, 这严重影响了系统的可用性以及用户搜索的效率.

1 方法概述

本文提出的语义网文档搜索方法的交互过程类似于传统的 Web 搜索引擎, 用户给定若干关键词作为查询, 系统返回匹配该查询的语义网文档集合. 该方法主要特点在于两个方面: 语义网文档模型和搜索结果的呈现.

在文档模型上, 采用的是基于 RDF 句子[5] 的向量空间模型, 并引入 URI 资源的权威描述的概念. 整个语义网文档的索引过程如图 3 所示. 给定一个语义网文档, 首先由[RDF Sentence Extractor]将其转换成 RDF 句子的集合; 得到的 RDF 句子集合分两步进行处理, 最终都交给索引组件[Indexer]构建索引. 第 1 步如图中的左半部分所示, 词向量生成器[sentence text vector generator]结合 URI 资源的权威描述索引[URI AuthLab Index]为 RDF 句子生成相应的词向量交给[Indexer]. 第 2 步如图中右半部

分所示, 文档元数据提取器[metadata extractor] 从句子集合中提取出元数据提交给[Indexer] 构建索引. 最后[Indexer] 将上述两步的处理结果生成最终的倒排索引[extended inverted index].

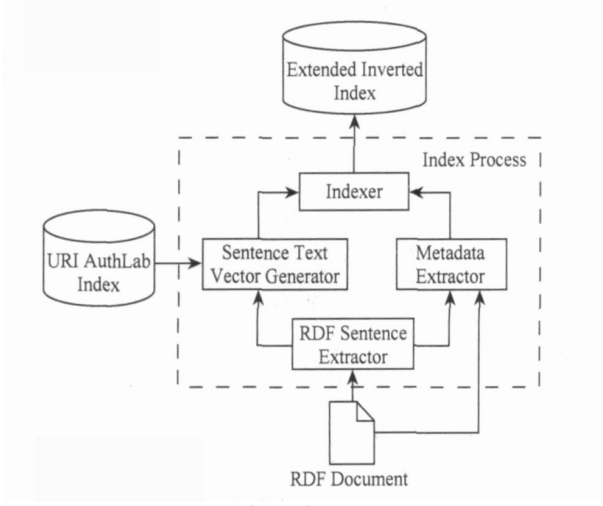

Fig. 3 The indexing process.
图 3 索引结构图

结果的呈现方面如图 4 所示, 采用基于 RDF 句子的内容提取机制, 包括两方面的信息: 1) 文档元数据. 从语义网文档中提取出文档元数据, 从中筛选出能够帮助用户快速理解该文档的元数据信息显示给用户. 2) 匹配证据的显示. 在结果页面中, 从每个匹配查询的文档中提取出包含用户查询关键词的最小 RDF 句子集合, 作为证据显示给用户. 这两种信息分别从文档本身的描述以及查询相关的角度帮助用户从结果列表中快速定位到需要的文档.

Fig. 4 The user interface of Falcons document search.
图 4 Falcons 文档搜索服务界面①

2 语义网文档的词向量

如上所述, 为了更加准确地表达语义网文档的语义, 需要一种机制根据 RDF 数据模型的特征将语义网文档转换为相应的词向量. 这样的向量需要满足以下两个需求: 首先, 需要支持基于 URI 重用的数据共享; 其次, 需要保留 RDF 数据的语义信息. 本文采用将 RDF 句子作为最小语义单元的方案, 在 RDF 句子的基础上构建词向量.

2.1 URI 的权威描述

RDF 模型使用 URI 指称对象和概念. 由于这种 URI 指代的存在会使得仅从当前文档的信息来生成相应的词向量导致信息缺失, 从而无法很好地满足用户的搜索需求. 在自然语言中, 要解决指代引起的信息缺失问题一个简单而有效的方案是将代词替换成其所指代的名词. 类似地, 对于 URI 的指代, 可以找出被指代资源的描述信息来进行替换. 从搜索角度来看, 这种信息需要满足下面 3 个要求: 一方面, 通过这个描述信息能够了解这个 URI 指称的是什么资源; 另一方面, 这个描述信息的范围必须严格控制, 避免引入太多的信息, 最终影响搜索系统的精度, 比如一个人的简介信息通常会包含很多其他信息, 就不适合全部引入; 最后, 描述信息的数据来源也是需要考虑的. 对语义网中实际数据的分析表明, 一个 URI 经常在不同的文档中进行重复定义或描述(作为三元组的主语) . 这种有意或无意的重新定义通常是不一致性数据或者错误数据产生的主要来源. 因此, 本文仅考虑 URI 的权威描述. 根据文献[6], 一个 URI 的权威描述是指其所隶属的域的拥有者的对其进行的描述.

综合上述讨论, 提出 URI 权威描述的概念来解决 URI 指代问题. 定义 1 给出了 URI 权威描述的形式化定义. 权威描述表示为词向量的形式, 其中 β_1 和 β_2 表示在标签词向量和本地名称词向量之间的权重分配.

定义 **1.** URI 的权威描述. 给定一个 URI 资源 u, 它的权威描述可以定义为

$$\boldsymbol{Auth\_Label}(u) = \beta_1 \times \boldsymbol{AuthLab}(u) + \beta_2 \times \boldsymbol{LN}(u),$$

其中, $\boldsymbol{AuthLab}(u)$ 是 u 的权威语义网文档中定义的所有标签($rdfs:label$)的词向量; $\boldsymbol{LN}(u)$ 是其本地名称(local name) 的词向量.

2.2 RDF 句子的词向量

除了使用 URI 来指称资源之外, RDF 数据模型的另外一个重要特征是采用三元组的形式来陈述命题. RDF 是一种基于图的数据模型, 其抽象句法

① http://iws.seu.edu.cn/services/falcons/documentsearch/index.jsp

是三元组的集合, 称之为 RDF Graph[7]. 对于图 1 中的 RDF 数据片段而言, 其对应的 RDF Graph 可表示为图 5 所示的结构.

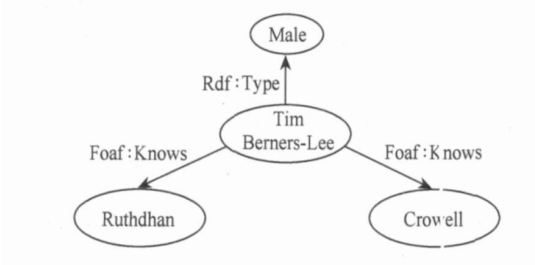

Fig. 5 An example of RDF graph.

图 5 RDF Graph 示例

语义网文档是对其所描述的 RDF Graph 的一种序列化表示, 这种序列化采用某种语法, 如 RDF/XML, N-Triples 或者 N3 对 RDF Graph 进行编码, 其目的是以文本的方式保留 RDF Graph 的结构信息, 以便传播和重建相应的 RDF Graph. 它并不能保证生成的语义网文档中词的词频能相应地反映 RDF Graph 的语义. 例如, 图 1 是采用 RDF/XML 的语法对图 5 中的 RDF Graph 的一种序列化表示, 从它的词频来看, Ruthdhan 和 Crowell 的词频都高于 Berners, 而图 5 很直观地反映了这段数据更多描述的是 Tim Berners-Lee 的信息. 因此, 如果采用向量空间的模型, 那么就需要一种词向量表示方法, 能够将 RDF Graph 的语义更准确地反映到相应的词向量中.

RDF Graph 是 RDF 三元组的集合. 一种可行的方案就是为每条三元组生成词向量, 然后综合生成整个 RDF Graph 的词向量. 在这种方式下, RDF Graph 中的节点的对应描述将依照其度数(即相关三元组的个数) 计入到最终的词向量中, 从而最终的词频能够更好地反映 RDF Graph 的语义. 例如图 5 中, Tim Berners-Lee 的描述将计数 3 次, 其他两人的描述都是 1 次, 所以这种方法更加准确地反映了该段数据描述的语义.

准确地体现文档语义是词向量构建的基本要求, 除此之外, 词向量的构建还需要能够记录词在语义网文档中出现的语义相对完整的语义单元, 这样才能提供更加适合理解的关键词匹配的证据(下节讨论). 然而基于 RDF 三元组的方案并不能满足这个要求. 这是由于 RDF 规范中引入了空白节点的概念. 共享空白节点的最小 RDF 三元组的集合构成一个原子的语义结构, 即语义上不可分割. 文献[5]提出 RDF 句子的概念来解决 RDF 三元组的语义可能不完整的问题. RDF 句子是三元组的集合, 其特性是能够确保一个原子的语义结构不被破坏. 因此, 本文将语义网文档看作 RDF 句子的集合, 以 RDF 句子代替 RDF 三元组作为最小的语义单元来生成语义网文档的词向量.

定义 **2.** RDF 句子的词向量. RDF 句子 $\bar{s}$ 的词向量定义为

$$\boldsymbol{Text\_Desc}(\bar{s}) = \sum_{s,p,o \in \bar{s}} \alpha_1 \cdot \boldsymbol{Text}(s) + \alpha_2 \cdot \boldsymbol{Text}(p) + \alpha_3 \cdot \boldsymbol{Text}(o);$$

$$\forall r \in \{s\} \cup \{p\} \cup \{o\},$$

$$\boldsymbol{Text}(r) = \begin{cases} \boldsymbol{Auth\_Label}(r), & r \in U; \\ \boldsymbol{LN}(r), & r \in B; \\ \boldsymbol{LF}(r), & r \in L; \end{cases}$$

其中, U, B, L 分别是 *URI*、空白节点以及字面量的集合; $\boldsymbol{LF}(r)$ 是字面量文本形式的词向量.

定义 2 给出了 RDF 句子的词向量的形式化定义. 词向量是由 RDF 句子包含的三元组的词向量组合而得. 有了 RDF 句子的词向量, 一个语义网文档的词向量就是该文档所包含的 RDF 句子的词向量之和(见定义 3). 从定义 2 和定义 3 可以看出, 基于 RDF 句子的词向量具有以下两种特性: 首先, 语义网文档的最终词向量仍然是 RDF 三元组的向量之和, 因此, 向量的词频能够更好地反映语义网文档的语义信息; 其次, 系统能够记录每个词所出现的相对完整的语义单元(即 RDF 句子) 的列表. 为了记录这样的位置信息, 我们对传统的倒排索引进行了扩展. 限于文章篇幅, 本文不讨论扩展索引的结构以及相关的技术.

定义 **3.** 语义网文档的词向量. 给定一个语义网文档 d, 令 $Sent(d)$ 为从 d 中提取的所有 RDF 句子的集合. d 的词向量可定义为

$$\boldsymbol{Text\_Desc}(d) = \sum_{\bar{s} \in Sent(d)} \boldsymbol{Text\_Desc}(\bar{s}).$$

3 结果文档的内容片段提取

对于搜索引擎而言, 从结果文档中提取查询相关的片段对于系统的可用性至关重要. 引言中的分析指出语义网文档本身的特性使得传统 Web 搜索引擎的提取方法不再适用. 本节提出一种基于 RDF 句子的提取方法: 以 RDF 句子作为最小的语义单元, 从匹配的文档中提取语义紧凑的句子集合作为证据显示给用户.

查询证据的定义: 一个关键词匹配的 RDF 句子

是指词向量中包含此关键词的 RDF 句子. 定义 4 给出了关键词匹配的 RDF 句子的形式化定义. 一个查询可能包含多个关键词, 在这种情况下会有多个匹配查询的 RDF 句子的列表. 如何从中选出适合的子集作为查询匹配的证据是需要考虑的问题. 本文考虑的原则是从匹配的 RDF 句子列表中选出包含所有关键词的最小子集. 因为对于多关键词查询而言, 关键词之间通常存在某种语义上的关联. 如果证据中 RDF 句子个数越少通常其语义就越紧凑, 也就更加能够反映查询关键词之间的关联. 根据这个原则, 定义 5 给出了关键词查询的证据的严格定义.

定义 **4.** 关键词匹配的 RDF 句子. 设 d 为一个语义网文档, k 为一个关键词, $\boldsymbol{V}_k$ 为 k 在向量空间中相应的向量表示, 则 d 中匹配 k 的 RDF 句子集合为

$$S_k(d) = \{\tilde{s} \mid \boldsymbol{Text}(\tilde{s}) \cdot \boldsymbol{V}_k \neq 0\}.$$

定义 **5.** 关键词查询的证据. 设 $q=\{k_1, k_2, \cdots, k_n\}$ 为一个查询, d 为一个语义网文档, 则有:

$$\hat{S}_q(d) = \{\tilde{S} \subseteq Sent(d) \mid \forall k_i \in q, \tilde{S} \cap S_{k_i}(d) \neq \varnothing\}.$$

进而, 文档 d 匹配查询 q 的证据集:

$$E_q(d) = \arg \min_{\tilde{S} \in \hat{S}_q(d)} \left(\mid \tilde{S} \mid \right).$$

$\forall e_q(d) \in E_q(d)$, $e_q(d)$ 为 q 匹配文档 d 的一个证据.

可以证明从文档(RDF 句子集合)中寻找定义 5 给出的查询证据的问题是集合覆盖问题. 集合覆盖问题已经被证明是 NP 完全问题[8]. 在诸多近似算法中贪心算法实现最简单, 且执行效率很高[9]. 本文采用基于贪心算法的证据寻找算法, 限于篇幅, 略去算法的细节和算法复杂性讨论.

4 相关工作

在语义网、XML 以及数据库领域, 有一系列的工作通过扩展向量空间模型来解决结构化数据的关键词搜索问题. 根据搜索的对象不同它们大致分为两类: 文档搜索和数据搜索.

在语义网文档搜索领域文献[10]记录每个词出现的位置, 将关键词的距离信息引入到匹配模型中以提高搜索精度, 然而该方法在词频上并没有考虑 RDF 图模型本身的结构信息. 在 XML 文档搜索领域[11-12]将向量空间每个维度扩展为(t, c)对, 其中 t 是词汇, c 是语境, 这种扩展的目的是支持弱结构化的查询. 这类扩展是基于 XML 文档的树形结构, 而本文是基于 RDF 数据的抽象语法 RDF 图对向量空间进行扩展.

对于数据搜索, 其搜索目标是从结构化数据中提取匹配关键词的子集作为答案返回. XML 搜索领域有文献[13-15], 数据库领域的代表论著有文献[16-18]. 这类方法应用场景是 Query/Answer 即直接给出问题的答案, 研究的重点是图(树)的索引技术. 本文重点研究文档的词向量的构建方法和查询证据的提取. 相似之处在于都需要从信息空间中提取子集, 本文的提取问题是集合覆盖问题, 而数据搜索相关工作是研究子图(子树)搜索问题.

5 实　　验

本节分别从搜索系统的精度、召回率以及可用性 3 个方面对本文提出的方法进行评估. 前两个实验采用的数据集是 Falcons 数据集中所有描述 FOAF 信息的文档集合. 共计 63941 个语义网文档, 包含 4438751 个 RDF 句子.

5.1 比较传统文档处理方法和基于 RDF 句子的方法(精度)

提出基于 RDF 句子的词向量构建方法, 其目的是更好地反映语义网文档的语义. 在搜索指标上期待的结果是该方法应该具有更好的精度表现. 评估搜索系统的精度通常需要 3 个要素: 数据集、一组查询以及在给定数据集上每个查询对应的正确答案即相关文档集合. 比如 Text REtrieval Conference (TREC) 就为传统信息检索系统的评估提供了丰富的数据集、Topics(查询主题)和 Topics 对应的正确答案.

语义网文档检索领域没有类似于 TREC 的公开数据集可以采用. 本文根据所采用的 FOAF 数据集的特征, 模拟出该数据集上的查询以及查询对应的正确答案. 方法是将 FOAF 文档作者的名字作为查询, 其 FOAF 文档作为正确答案. 按照这种方法辅以人工验证的方式从数据集中抽取了 387 个这样的 FOAF 文档, 生成相应的 387 个(查询, 相关文档)对. 这样我们就构建了基于 FOAF 数据集的测试集. 在 FOAF 数据集中对上述 387 个查询进行实验, 记录每次查询结果中相关文档获得的排名. 图 6(a)和图 6(b)给出了传统文档处理方法和基于 RDF 句子的方法的实验结果比较. 首先, 在 387 次查询中, 对相关文档排序值小于等于 k 的查询个数进行

计数, k 从 1~ 30. 从图 6(a) 可以看出, 基于 RDF 句子的方法要优于传统 IR 的技术方案. 图 6(b) 给出了两种方法前 k 个结果包含标准结果的可能性, 图中下方点划线样式的曲线是两种方法可能性之差. 当 k 取 1, 10 和 20 时, 两者相差百分比分别为 19. 12%, 7. 45% 和 9. 82%. k 取 1(即第 1 个文档是正确答案) 时, 两者的相差程度最大. 当 k 大于 3 后, 两者的相差稳定在 8%~ 10% 左右.

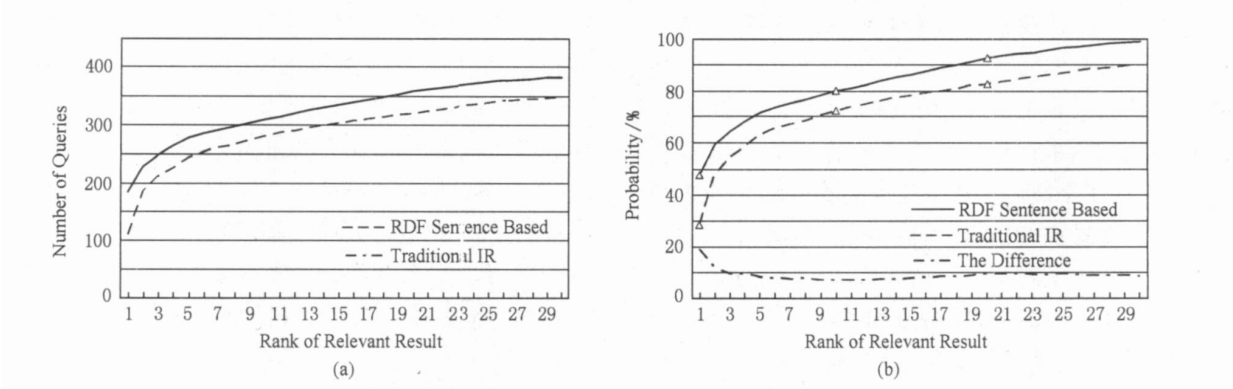

Fig. 6 The comparison of precisions between traditional IR and RDF sentence based approach. (a) Number of queries each of which the relevant document is in top-k results(k = {1, 2, 3, …, 30}) and (b) The probability of right answers in the top-k results.

图 6 传统 IR 与基于 RDF 句子方法的精度比较. (a) 相关文档在前 k 个结果内的查询个数(k = {1, 2, 3, …, 30}); (b) 前 k 个结果包含标准结果的可能性

5.2 引入被重用的 URI 的权威描述(召回率)

引入 URI 的权威描述能够解决 URI 指称带来的信息缺失问题. 从而提升搜索系统的召回率. 在 FOAF 数据集上构建了两个索引, 分别是引入和不引入 URI 的权威描述信息. 针对这两个索引, 执行了 50 个相同的查询. 结果显示在图 7 中, 横坐标代表每次查询, 纵坐标代表查询结果集的大小. 经统计, 在引入 URI 的权威描述之后, 有近 60% 的查询获得了更多的结果, 平均提升大约 29%.

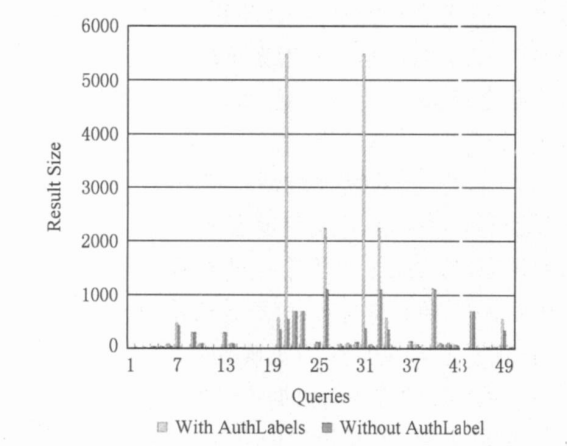

Fig. 7 The comparison of result sizes of 50 queries on different indices.

图 7 不同索引上 50 个查询的结果集大小对比

5.3 结果文档的内容提取对于系统可用性的提升

为了从可用性角度评估结果文档内容提取方案的有效性, 设计了以下实验.

实验由 4 个搜索任务组成, 每个任务针对一个搜索主题(主题取自 TREC robust test set). 实验参与者首先阅读当前主题的描述, 然后阅读给定的 3 个不同搜索引擎的第 1 个结果页面, 并记录能够确认相关或者不相关的文档个数. 该实验的假设是能够确定相关与否的文档比重, 比重越大说明内容片段的提取越合理, 系统的可用性也就越高.

共有 10 位志愿者参与此实验, 其中 8 位完成了实验. 实验中比较的 3 个搜索引擎分别是 Falcons, Sindice[2] 和 Watson[3]. 图 8 给出了实验结果. 其中每个搜索引擎名字后面的括号里的数字代表结果页中包含的文档总数, 如 Sindice(10) 表示 Sindice 第 1 页包括 10 个文档. 表中每行表示一个用户的实验记录: RL 表示相关, NRL 表示不相关. 分别记录的是用户确认相关和不相关的文档个数, Ratio 是比率, 表示的是(相关+ 不相关)/文档总数. 比率值是系统结果页面信息组织和呈现方案优劣度的指示符, 也是实验中我们最关心的数值. 最后一行给出平均比率值. 4 个主题的实验结果如图 8(a) ~ (d) 所示. 总体而言, Sindice, Falcons 和 Watson 三个系统的平均比率值分别为 25. 63%, 81. 61% 和 12. 89%. 可以看出, 在使用 Falcons 系统时, 实验者在绝大部分情况下可以明确地判断一个文档是否相关, 相比于其他语义网文档搜索引擎具有明显的优势. 从而使得语义网文档搜索系统在可用性上有了很大的提升.

User	Sindice(10)			Falcons(7)			Watson(3)		
	RL	NRL	Ratio/%	RL	NRL	Ratio/%	RL	NRL	Ratio/%
U1	1	0	10.00	6	1	100.00	0	0	0.00
U2	4	0	40.00	7	0	100.00	0	0	0.00
U3	2	4	60.00	4	1	71.43	0	3	100.00
U4	5	2	70.00	5	0	71.43	0	0	0.00
U5	3	1	40.00	5	0	71.43	0	0	0.00
U6	2	0	20.00	7	0	100.00	0	0	0.00
U7	2	0	20.00	7	0	100.00	0	0	0.00
U8	2	0	20.00	4	0	57.14	0	0	0.00
Avg.			35.00			83.93			12.50

(a)

User	Sindice(10)			Falcons(7)			Watson(3)		
	RL	NRL	Ratio/%	RL	NRL	Ratio/%	RL	NRL	Ratio/%
U1	2	0	20.00	8	1	90.00	0	0	0.00
U2	2	1	30.00	10	0	100.00	1	0	33.33
U3	1	1	20.00	2	4	60.00	2	1	100.00
U4	2	1	30.00	5	1	60.00	0	0	0.00
U5	3	1	40.00	4	3	70.00	0	0	0.00
U6	3	3	60.00	5	1	60.00	1	0	33.33
U7	2	0	20.00	5	3	80.00	1	0	33.33
U8	1	0	10.00	8	0	80.00	0	0	0.00
Avg.			28.75			75.00			25.00

(b)

User	Sindice(10)			Falcons(7)			Watson(3)		
	RL	NRL	Ratio/%	RL	NRL	Ratio/%	RL	NRL	Ratio/%
U1	2	0	20.00	5	4	90.00	0	0	0.00
U2	2	0	20.00	10	0	100.00	0	0	0.00
U3	1	4	50.00	5	2	70.00	0	0	0.00
U4	3	0	30.00	8	2	100.00	5	0	62.50
U5	2	0	20.00	8	2	100.00	0	0	0.00
U6	2	0	20.00	7	0	70.00	0	0	0.00
U7	2	0	20.00	5	5	100.00	0	0	0.00
U8	3	0	30.00	7	1	80.00	0	0	0.00
Avg.			26.25			88.75			7.81

(c)

User	Sindice(10)			Falcons(7)			Watson(3)		
	RL	NRL	Ratio/%	RL	NRL	Ratio/%	RL	NRL	Ratio/%
U1	1	0	10.00	7	1	80.00	0	0	0.00
U2	2	0	20.00	9	0	90.00	0	0	0.00
U3	1	0	10.00	5	0	50.00	0	0	0.00
U4	2	0	20.00	10	0	100.00	3	1	40.00
U5	1	0	10.00	5	1	60.00	0	0	0.00
U6	1	0	10.00	8	0	80.00	0	0	0.00
U7	1	0	10.00	7	3	100.00	0	0	0.00
U8	1	0	10.00	6	1	70.00	1	0	10.00
Avg.			12.50			78.75			6.25

(d)

Fig. 8 The ratio of relevance-decidable documents in 4 search topics. (a) Magnetic levitation; (b) Implant dentistry; (c) Industrial espionage; and (d) Journalist risks.

图 8 4 个主题搜索结果中用户能够确认相关与否的文档数与结果总数的比率. (a) 磁力悬浮; (b) 口腔种植/人工牙根; (c) 工业间谍活动; (d) 新闻记者的危险

6 总 结

本文给出了基于RDF句子的语义网文档的词向量构建方法. 该方法很好地解决了在向量空间模型下直接对语义网文档生成向量的方法所导致的丢失结构信息的问题. 在文档向量中引入了资源的权威描述, 使得系统能够搜索到跨文档互连的RDF数据. 在生成结果页面时, 以RDF句子为单位, 从结果文档中提取文档元数据信息和查询相关的证据信息, 使得用户在绝大部分情况下能够快速确认一个文档是否相关. 实验表明, 这种方法在精度、召回率以及可用性方面都有较好的表现. 在精度方面, Top1, Top10以及Top20 3个粒度分别有大约19%, 7%以及10%的提升. 在召回率上, 实验给出的50个查询中有60%获得了提升, 平均提升29%左右. 最后, 在系统可用性的比较中(在本文给出的指标上)超出其他语义网文档搜索引擎近32%.

参 考 文 献

[1] Ding L, et al. Swoogle: A search and metadata engine for the semantic Web [C] //Proc of the 13th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2004: 652-659

[2] Tummarello G, Delbru R, Sindice E O. Com: Weaving the open linked data [C] //Proc of the 6th Int and 2nd Asian Semantic Web Conference (ISWC2007 + ASWC2007). Berlin: Springer, 2007: 552-565

[3] Watson-d'Aquin M, et al. WATSON: A gateway for the semantic Web [C] //Proc of European Semantic Web Conference 2007. Berlin: Springer, 2007

[4] Ding L, et al. Finding and ranking knowledge on the semantic Web [C] //Proc of the 4th Int Semantic Web Conference(ISWC 2005). Berlin: Springer, 2005: 156-170

[5] Zhang X, Cheng G, Qu Y. Ontology summarization based on RDF sentence graph [C] //Proc of the 16th Int Conf on World Wide Web. New York: ACM, 2007: 707-716

[6] Jacobs I, Walsh N. Architecture of the World Wide Web [EB/OL]. (2004-12-01) [2009-06-05]. http://www.w3.org/TR/webarch

[7] Klyne G, Carroll J J. Resource Description Framework (RDF): Concepts and Abstract Syntax [EB/OL]. (2004-02-01) [2009-06-05]. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210

[8] Mannino C, Sassano A. Solving hard set covering problems [J]. Operations Research Letters, 1995, 18(1): 1-5

[9] Chen Ruibing, Huang Wenqi. A heuristic algorithm for set covering problem [J]. Computer Science, 2007, 34(4): 133-136
(陈端兵, 黄文奇. 一种求解集合覆盖问题的启发式算法 [J]. 计算机科学, 2007, 34(4): 133-136)

[10] Wu Gang, et al. Fine-grained semantic Web retrieval [J]. Journal of Tsinghua University: Science and Technology, 2005, 45(S1): 1865-1872 (in Chinese)
(吴刚, 等. 细粒度语义网检索[J]. 清华大学学报: 自然科学版, 2005, 45(S1): 1865-1872)

[11] Carmel D, et al. Searching XML documents via XML fragments [C] //Proc of the 26th Annual Int ACM SIGIR Conf on Research and Development in Informaion Retrieval. New York: ACM, 2003: 151-158

[12] Chu-Carroll J, et al. Semantic search via XML fragments: A high-precision approach to IR [C] //Proc of the 29th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2006: 445-452

[13] Cohen S, et al. XSEarch: A semantic search engine for XML [C] //Proc of the 29th Int Conf on Very Large Data Bases. San Francisco: Morgan Kaufmann, 2003: 45-56

[14] Guo L, et al. XRANK: Ranked keyword search over XML documents [C] //Proc of the 2003 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2003: 16-27

[15] Li Y, Yu C, Jagadish H V. Schema-free query [C] //Proc of the 30th Int Conf on Very Large Data Bases. San Francisco: Morgan Kaufmann, 2004: 72-83

[16] Hristidis, Gravano V L, Papakonstantinou Y. Efficient IR-style keyword search over relational databases [C] //Proc of the 29th Int Conf on Very Large Data Bases. San Francisco: Morgan Kaufmann, 2003: 850-861

[17] Kacholia V, et al. Bidirectional expansion for keyword search on graph databases [C] //Proc of the 31st Int Conf on Very Large Data Bases. New York: ACM, 2005: 505-516

[18] He H, et al. BLINKS: Ranked keyword searches on graphs [C] //Proc of the 2007 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2007: 305-316

Wu Honghan, born in 1976. PhD candidate in the School of Computer Science and Engineering, the Southeast University, China. He received his BS degree in mechanical engineering from the Southeast University in 1998. His main research interests include semantic Web, data pattern analysis and semantic Web search.

吴鸿汉, 1976年生, 博士研究生, 主要研究方向为语义Web、数据模式分析以及语义网搜索问题.

Qu Yuzhong, born in 1965. Professor and PhD supervisor of the School of Computer Science and Engineering, the Southeast University. His main research interests include software engineering, semantic Web and Internet computing.

瞿裕忠, 1965 年生, 教授, 博士生导师, 主要研究方向为软件方法与技术、语义 Web 和万维网科学.

Li Huiying, born in 1977. PhD candidate and lecturer of the School of Computer Science and Engineering, the Southeast University. Her main research interests include semantic Web search and data management.

李慧颖, 1977 年生, 博士研究生, 讲师, 主要研究方向为语义 Web 搜索和语义网数据管理.

Research Background

Semantic Web document search is one of the most efficient ways to find and reuse semantic Web data. Most of existing approaches are based on traditional information retrieval technologies. However, the characteristics of RDF data model make semantic Web documents be inherently different from traditional natural language oriented documents. We identified that traditional IR technologies are short in preserving RDF data structures, processing linked data across documents and generating reasonable snippets. To address these obstacles we proposed an RDF sentence based approach. There are two key points. First, the document vectors are constructed based on the abstract syntax of RDF documents—RDF graph instead of a literal interpretation. Secondly, the authoritative description of URI resources is introduced. When constructing document vectors, these descriptions of reused URIs can be included and this makes linked data searchable. Our work is supported by the National Natural Science Foundation of China (60773106) and Natural Science Foundation of Jiangsu Province (BK2008290).

第 17 届全国网络与数据通信学术会议(NDCC2010)征文通知

2010 年 9 月 16 日—17 日　北戴河

由中国计算机学会网络与数据通信专业委员会主办, 由东北大学秦皇岛分校和东北大学信息科学与工程学院联合承办的"第 17 届全国网络与数据通信学术会议"将于 2010 年 9 月 16 日到 17 日在美丽的海滨城市北戴河举行.

本次大会将围绕"网络与通信新技术及应用"这一主题展开, 为来自国内外高等院校、科研院所、企事单位的学者、教授、专家、工程师提供一个代表国内网络与数据通信产学研界高水平的高层信息交流平台, 探讨本领域发展所面临的关键挑战问题和热点研究方向.

会议论文集将由《东北大学学报(自然科学版)》增刊出版(EI 检索源). 论文参照《东北大学学报》格式, 字数一般不超过 6000 字, 稿件通过投稿系统提交, 具体参见会议网站 http://ndcc2010.neuq.edu.cn. 部分优秀论文将被推荐到《计算机学报》、《电子学报》、《计算机研究与发展》、《电子与信息学报》的正刊发表(均为 EI 检索源). 会议期间将评选会议优秀论文和优秀学生论文.

本次会议的主要征文范围包括以下领域(但不限于):

新一代网络技术: 网络体系结构、路由/交换技术、协议工程、网络虚拟化、认知网络、IPv4/IPv6 过渡技术、NGN/NGI 平台应用; 新一代计算技术: 云计算/网格计算、并行/分布式计算、普适/效用计算、服务计算; 无线通信技术: 下一代移动通信技术、自适应信号处理、传感器网络、移动自组织网络、智能天线、卫星通信; 其他: 光通信技术、网络安全、网络管理、网络应用.

投稿须知

1) 投稿内容突出作者的创新与成果, 具有较重要的学术价值与应用推广价值, 未在国内外公开发行的刊物或会议上发表或宣读过.

2) 论文语言要求中文, 字数一般不超过 6000 字, 论文格式参照《东北大学学报》, 投稿稿件用 Word 文件形式. 东北大学学报的网站地址如下: http://xuebao.neu.edu.cn/natural/index.asp.

3) 请在稿件最后附上第一作者姓名、性别、职务/职称、所属单位、通信地址、邮政编码、联系电话和 E-mail 地址, 并注明论文所属领域.

4) 被录用的论文, 至少要有一位作者参加会议并发言, 才有资格参与优秀论文的评选.

投稿方式

论文投稿通过投稿系统进行提交, 详见会议网站: http://ndcc2010.neuq.edu.cn 或者通过邮件地址 ndcc2010@mail.neuq.edu.cn 联系. 电子邮件请在邮件标题注明"NDCC2010 投稿".

重要日期

论文提交截止日期: 2010 年 5 月 15 日; 论文录用通知日期: 2010 年 7 月 1 日; 会议注册截止日期: 2010 年 8 月 15 日.

联系方式

联系电话: 0335-8052155　　联系人: 王翠荣　韩来权

邮件地址: ndcc2010@mail.neuq.edu.cn　　会议网站: http://ndcc2010.neuq.edu.cn