

一种有效的基于本体的词语-概念映射方法

李文 陈叶旺 彭鑫 赵文耘

(复旦大学计算机科学技术学院, 上海 200433)

摘要: 词语-概念映射是实现语义检索的重要一环, 对语义检索结果有很大的影响. 提出了一种词语-概念映射方法, 该方法利用概念-文档与词语-文档两重关系, 计算出词语与概念的相关度与置信度, 再实现词语(集)映射到本体概念(集)。从实验结果来看, 有效地提高了检索的效果。

关键词 本体, 词语, 映射, 标注, 三元组

An Effective Term-Concept Mapping Method Based on Ontology

LI Wen CHENG Yewang PENG Xin ZHAO Wenyun

(School of Computer Science, Fudan University, Shanghai 200433, China)

Abstract: As one of the key steps of ontology-based semantic search, term-concept mapping has a great impact on precision and recall. In this paper, we propose a novel term-concept mapping method, which takes into count both the concept-documents and term-document, calculates the correlation and confident degree between them firstly, and then gives the final mapping results. The experiments show that this method improves the precision and recall effectively.

Key words ontology, term, mapping, annotation, triple

1 引言

基于语义的查询, 作为从上世纪 90 年代问世的语义网^[3]在信息检索领域的一个应用, 已经在改进检索查全率和查准率方面表现出巨大的潜力。与传统的基于关键词及词频的信息检索引擎相比, 语义检索引擎通过在文本中加入语义 tag, 这种通过语义 tag 结构化, 概念化文档中的对象的方法, 比传统基于关键词搜索的方法在挖掘用户给出信息中隐含的信息方面更加有效^[1]。

在基于本体^[3]的语义检索系统中, 通常使用本体作为描述领域知识的基础, 将文档标注到相关的本体知识概念中。在检索时, 通过将关键词匹配到本体知识概念, 并根据标注的信息以找到相关文档。然而对于所有信息检索系统来说, 存在着一个根本问题: 关键词错误匹配^[12], 即系统的开发人员与用户使用不同的术语来描述相同的概念, 如单车与自行车描述的是同一事物。这本质上是由于人类对于世界的不同理解所造成的, 同时也是不可避免的, 这对于系统的查全率有着很大的影响。此外, 从用户对信息检索系统的使用情况来看, 查询语句变得越来越简短也使得词语错误匹配的问题变得更加突出, 因此无论是传统基于关键词匹配或者是基于语义的查询, 在接受用户以自然语言(如关键词, 疑问句)方式提交的查询之后, 通常都使用查询扩展来提高查全率, 因为通过查询扩展^[2]可以利用与查询关键词相关的词语对查询进行修正以找到更多, 更准确的相关文档。

而在查询扩展中, 一个关键的问题是如何找到与查询关键词相匹配的概念集。在基于本体的检索系统中, 这个问题就转化为词语-本体知识之间的映射。词汇-本体知识映射在本体和

语义网领域中有着大量的研究工作, 例如^[16]使用相似度计算, 在两个机器系统之间自动完成本体概念的映射。总的来说, 现有的方法基本上是基于关键词方式来完成词语-知识映射的, 这种方法常常会带来许多语义理解错误,^[15]中称其为词语问题(vocabulary problems)。如, 同义词问题(synonyms)、歧义问题(polysemy)、异体问题(lemmas)、准同义问题(quasi-synonyms)等, 在提高查全率的同时难以保证查准率。产生词语问题的根本原因在于, 人们在现实生活中描述同样的对象或事件的用词存在着多样性。在基于本体的语义检索系统中, 目前大多数利用文档对本体标注都是以单个本体知识-单个文档的形式进行的, 虽然当中可以有对多的关系, 但我们认为这种方法丢掉了语义的信息。因此, 在我们的方法中, 考虑将文档标注以文档-本体关系来进行。本文提出了一种基于词语-概念三元组-文档标注结合的方法来计算词语-本体知识关系的相关度。该方法扩展了传统的单个词语到单个概念的方法, 通过将词语(集)映射到本体中的知识关系, 考虑上下文信息, 以计算词语与三元组中对应知识的相关度。

文章的结构如下: 第二节介绍了已有的相关工作, 第三节中概述了三元组(链)-文档标注方法, 第四节介绍基于三元组-文档标注的词语-本体知识关系相关度计算, 第五节描述了方法的具体实现及其与有方法的对比, 第六节进行总结。

2 相关工作

随着语义网的大幅进展, 以及人们所面临的下一代搜索的挑战和语义万维网的兴起等, 极大地推动了语义搜索这一潜在研究领域在最近的发展, 并且促成了语义搜索这一名词的广泛

本文受国家 863 高技术研究发展计划基金项目(2007AA01Z179)资助

李文(1985-), 男, 硕士生。主要研究方向为语义网。Email: 072021123@fudan.edu.cn; 陈叶旺(1978-), 男, 博士生, 主要研究方向为语义网, 本体论等; 彭鑫(1979-), 男, 讲师, 主要研究方向包括构件技术与软件体系结构、软件产品线、软件维护与再工程等; 赵文耘(1964-), 男, 教授, 主要研究方向为软件工程、基于构件的软件开发方法, 软件开发工具及其环境, 企业应用集成(EAI)。

使用。最近几年,明确提出或提到语义搜索的研究工作越来越多的涌现出来^[4-7]。这些工作从各自不同领域的不同角度出發,以各种方式对各种不同类型的语义信息进行利用来提高搜索的效果。

目前已经有一些关于语义标注的研究工作,包括应用信息抽取技术的方法(Information Extraction, IE)、基于本体的信息抽取方法(Ontology Based Information extraction, OBIE)、基于自然语言处理的方法(Natural Language Processing, NLP)等。现将这些方法简单介绍如下:

1)应用传统信息抽取技术的方法。一些研究者应用传统信息抽取技术对网页进行语义标注。例如Amileare^[17]应用机器学习的方法在标注好的训练集上进行训练。通过提供不同领域的标注训练文档,它可以适应多个领域。S-Cream 综合运用了Ont-0-Mat^[18]与Amilcare 工具,将Amilcare产生的标注结果与Ont-0-Mat定义的关系元数据内的概念标记匹配。但是由于二者之间表示的差异,这种匹配十分困难,因此难以产生符合领域概念的标注结果。这种问题在应用传统信息抽取技术进行语义标注的方法中具有普遍性。传统的信息抽取技术不能有效地利用本体中已有的信息,而且IE抽取出的概念类型和关系与本体内部定义的概念和属性并不相符。在实际工作中,仍然需要额外定义匹配规则或使用其他方法来将IE结果转化为本体内知识。基于熵分类器语义角色标注^[13]是浅层语义分析的一种可行方案。描述了一个采用最大熵分类器的语义角色标注系统,把句法成分作为语义标注的基本单元,用最大熵分类器对句子中谓词的语义角色同时进行识别和分类,使用了一些有用的特征及其组合。在后处理阶段,在具有嵌套关系的结果中,只有概率最高的语义角色被保留。在预测了全部能够在句法分析树中找到匹配成分的角色以后,采用简单的后处理规则去识别那些找不到匹配成分的角色。

2)基于本体的信息抽取方法(OBIE)方法将本体作为信息抽取过程中可用资源的一部分,比如利用本体内已有实例信息来构造列表,简化抽取过程中对于概念实例的识别。例如 SemTag 首先在TAP本体实例集合中查找所有与待标注词匹配的可能实例集合,然后根据待标注词的上下文与实例集合中每个实例的上下文,分别构造各自的文本向量,进行相似度计算,找到与待标注词最匹配的实例。而KIM^[20]则可以在应用现有本体实例集对文本进行标注的同时,进行新实例的生成工作。OBIE方法的共性问题在于关注的是对文本中实例的标注与提取,却未能很好地抽取出实例间的关系。^[21]提出了一种有效的服务资源自动语义标注方法该方法将服务语义标注过程分解为域标注和概念标注两个阶段,重点针对域标注问题提出了基于K-NN的域标注算法,并通过对实际服务资源的域标注实验,验证了该算法的有效性。

3)基于自然语言处理的方法。为了处理自然语言超文本数

据,自然语言处理方法也是语义标注研究中一个重要的候选。文献^[6-8]中方法在完成句子中实例识别工作后,都注意到了分析自然语句过程中句法结构对语义标注的作用,并都试图从句子的主谓宾语法成分找到对应的 RDF 陈述,即从主语和宾语中找出对应领域本体概念的实例,然后将谓语动词映射为本体中属性。从谓语动词的词义中找到与其一致的领域本体属性是研究中的难点,而借助外部知识来辅助完成这一映射成为此类方法中的共同思想。其中 Artquakt^[9]和 i0ka^[10],都采用了依赖于通用语言本体来完成谓语动词语义与本体属性关系映射的方法。i0ka 框架处理中文自然语言文档,该框架借助 NLP 技术和本体定义可以从台湾股票领域的新闻网页中获取相关的知识。与文献^[6,7]方法不同,RelEx^[8]方法注意到了动词在表达关系时的作用及在本体构建时极少被用于属性定义的现状,采用了统计方法获取领域中表达属性关系的动词,并在句法分析过程中利用这些已标注类型的动词信息来寻找对应的本体属性。但此方法只关注了动词的重要性,却忽视了其他可能表达属性的词汇。此外,^[11]的方法在完成网页内实例识别后,作者提出了最小子树概念,利用 DOM 树结构来辅助关系抽取。

对于基于本体的语义搜索,词语到本体知识的映射是搜索关键步骤之一,匹配成功与否直接影响到最终的查准率和查全率。不同用户之间,用户与开发者之间对相同概念的理解偏差和用户无法准确地提供所需的信息是影响词语-本体知识关系匹配的两个重要因素。常用的计算相关度的方法有余弦相似度, Dice 相似度等。词语共现性是计算相关度的重要依据。通常来说,经常同时出现的词语相关度一般比较大。而计算共现性时,比较通用的依据有词语粒度,短语粒度^[17],概念粒度^[18,19]等等。由 Xu 和 Croft 提出的 LCA^[17]方法是一种经典的语义查询扩展方法,它最主要的贡献在于通过计算初始检索结果集的 top-k 文档中词语和查询中词语的共现度,以找到查询相关词语作为扩展词语。此外,计算相关度时还有一些其他重要因素,如句法上下文^[20],信息熵^[21]等。^[22]中提出了一个语义构件检索的理想场景,即通过语义推理和一系列问答式的会话过程,将一个本体中的概念映射到另一个本体中的概念。而^[10]中提出一种通过用户与构件库系统之间的会话式过程完成这一本体协商任务的方法,使用对动宾短语中谓语和宾语的联合匹配结合状语集合及定语集合在谓语和宾语刻画面属于词汇集合的概率进行映射。

3 三元组-文档标注方法

目前的语义标注工作,大部份还都只是用单个本体的 individual 为文档作标注,通过在文档中查找是否存在与 individual 相关的词汇来实现。我们认为这种方式割裂了 individual 所存在的语义环境,使得资源标注的结果不够准确,因为一个概念并不是单独存在的,它必须与其相关的各种属性

及属性值共存,才能体现出其语义,否则与普通的文本就没有区别了。我们将用单个本体的 individual 对文档进行标注的方法的不足归纳为以下三点:

(1)有些词汇意思相同,却常有不同的表达词组,如果只用 individual 标签值来做统计,往往会造成统计结果不准确(缺失了 individual 的各个属性值)。

(2)在含有英文的文本值中,单词的单复数形式、动词的过去式、动词完成式等等也常使统计结果不准确。

(3)最重要的是,这种方式割离了 individual 的存在语义环境,会使标注的结果可能完全错误。

由此我们提出,语义标注方式应该是基于本体语义关系,即本体三元组或三元组链的形式,亦即相当于用本体语义图中的子图来为资源作标注,不应当只是通过某个单独的本体知识来进行标注。

因而,我们之前的工作^[14]采用非内嵌方式来标注文档。从不同角度上来看,我们的方法属于半自动化方法。标注的结果存储在数据库中,而不是创建一个标注本体。每条标注都至少包含三方面内容,一个是本体知识关系,以三元组表示,第二个就是资源,包括资源标识和地址,第三个是前二者之间的相关度。当然,一个本体知识关系可以标注多个资源,一个资源可以被多个关系所标注,即二者之间可以是多对多的关系。例如,如表 1 为一个标注示例。

表 1: 语义标注结果

| 知识关系 | 资源 | 相关度 |
|---------------|-----------------------------|------|
| <黄瓜, 是, 植物> | www.fao.org/2009003/173.rdf | 0.81 |
| <白斑病, 危害, 黄瓜> | www.fao.org/2009003/173.rdf | 0.65 |
| <百菌清, 治, 白斑病> | 202.120.224.175/med/163.doc | 0.90 |
| ... | | |

为了方便叙述,我们进行以下假设约定: 设RB是一个信息资源库,包括各类文档,记为 $DS = \{d_1, d_2 \dots d_i \dots d_m\}$, 其中 d_i 是第 i 个文档, m 为文档数量, $0 < i \leq m$ 。

$T = \langle subj, pred, obj \rangle$ 表示一个三元组. 设TS是指用来标注资源的本体知识, 记为 $TS = \{t_1, t_2 \dots t_i \dots t_k\}$, 其中 t_i 是第 i 个三元组, k 为实例数量, $0 < i \leq k$ 。若一个文档 d 受本体知识关系 t 标注, 则记为 $r(t, d)$, 其中 d 表示一个文档资源, r 是前二者之间的相关度。

定义1 语义标注: 语义标注是从知识库文档库到标注结果的映射, 记为 $\delta: DS \times KB \rightarrow \{R(t, d, r)\}$, 其中 $0 < i \leq |DS|$, $0 < m \leq |TS|$ 。

4 映射方法

4.1 词语-概念相关度

在词语-标注文档-本体知识的相关关系中, 词语可能被包含在多个文档中, 而每个文档又可被一个或多个本体知识三元组标注, 通过统计包含词语的文档所属的本体三元组, 可以统计出这个词语对不同三元组的相关程度, 这种相关程度说明了词语-本体知识三元组间的联系紧密程度。

为计算这种相关关系, 我们通过以下规则, 来说明自然语言词汇对本体知识的相关关系。

规则 1: 如果词语集合通过文档映射到的三元组越多, 那么其对单个三元组的所属程度越低。

规则 1 是从三元组空间的分布来分析, 根据直觉, 如果词语集合与越多的三元组相关联, 说明词语集合对于这些三元组的区分度就越低, 与三元组的关联度也低。

规则 2: 如果词语集合通过越多的文档映射到同一个三元组, 那么词语集合对此三元组的所属程度越高。

规则 2 是从三元组的文档空间来分析, 如果词语集合在用三元组标注过的越多文档中出现, 说明这些词语在文档中分布越广越普遍, 因此与三元组的关联度也越高。

规则 3: 若词语集合通过文档映射到某个三元组, 如果三元组中对应本体实体标签词汇在词语集合中出现的个数越多, 那么词语集合与三元组的所属程度越高。

规则 3 是从三元组对应本体片段的标签词汇空间来分析的, 对于一个三元组 (有三个组成部分: [主 (subj), 谓 (pred), 宾 (obj)]) 当三个组成部分的标签词汇都出现在词语集合中时, 三元组与词语集合的相关度最高; 当有二个组成部分的标签词汇出现在词语集合中时, 相关度次之; 只有一个组成部分的标签词汇出现在词语集合中时, 相关度最低。

基于以上规则, 我们给出计算三元组 t 与词语集 c 的关联程度。

$$w(t, c) = \frac{tf(t, c) \times \log(n_t / N)}{\sqrt{\sum_{t \in c} [tf(t, c) \times \log(n_t / N)]^2}} \quad (1)$$

其中, $w(t, c)$ 为 t 与 c 的关联程度, 而 $tf(t, c)$ 为 t 在词语集 c 中的出现的次数, N 为 t 所标注过的词语集总数, n_t 为词语集中出现 t 的文本数, 分母为归一化因子。

规则 4: 令出现过词语集 c 的文档集合为 ds_c 。而知识关系三元组 t 标注的文档集合为 ds_t 。二者交集越大置信度越高, 二者并集越大置信度越低。

规则 4 是同时考虑词语集出现的文档集合与三元组标注文档集合。如果词语集出现过的文档集与知识关系三元组标注过的文档集的重合程度越大, 那么词语集与三元组的置信度也越高。

规则 4 是同时考虑词语集出现的文档集合与三元组标注文档集合。如果词语集出现过的文档集与知识关系三元组标注过的文档集的重合程度越大, 那么词语集与三元组的置信度也越高。

由此，我们给出词语集与三元组的置信度：

$$\tau(c, t) = \frac{ds_c \cap ds_t}{ds_c \cup ds_t} \quad (2)$$

其中 ds_c 表示出现过词语集中任一词语的文档集合， ds_t 表示三元组标注的文档集合。

由此，我们给出算法 1 计算考虑词语集合关联到三元组的文档空间以及三元组本身标注的文档空间之下，词语集合与三元组的置信度。于是，我们得到词语集 c 与三元组 t 的相关度

$$r(c, t) = w(c, t) \times \tau(c, t) \quad (3)$$

Algorithm 1

输入数据1: 与词语集合 d 相关的三元组及相应所属相关度二元组的集合 Γ

输出数据: 与词语集合 d 相关的三元组及相应置信度二元组得集合 Π

```
0 for each  $w(t, c)$  in  $\Gamma$ 
1   for each  $word$  in  $c$ 
2     for each  $doc$  in  $DS$ 
3       如果  $word$  在  $doc$  中的词频大于阈值且  $doc \notin ds_c$ 
4         那么 把  $doc$  加入到  $ds_c$  中
5       如果  $doc$  受  $t$  标注过
6         那么 把  $doc$  加入到  $ds_t$  中

7       计算  $\tau(c, t) = \frac{ds_w \cap ds_t}{ds_w \cup ds_t}$ 
8       将  $\tau$  加入到  $\Pi$ 
9 end
```

4.2 查询-文档相关度

计算查询词语与本体中概念相关度的目的是通过词语与概念的关系获取标注在本体概念之上的文档，以下我们给出查询-文档相关度的计算：

$$r(c, d) = r(c, t) \times r(t, d) \quad (4)$$

其中， $r(c, d)$ 是词语集与本体知识三元组的相关度， $r(t, d)$ 是对应本体三元组与文档的相关度。

5 实验及比较

5.1 实验设置

目前在语义搜索领域还没有一个公认的测试数据集和评价方法，因而，我们在做测试时，使用的数据是由我们自己根据过量粮农组织^[9] (FAO) 建立并公布的本体，抽取其中的部分建

立的本体知识，以及农业相关的资源，这些资源都经过我们的语义标注工具或手工方式标注过。本体知识使用的是农业病虫害本体 CropDisease (<http://icgr.caas.net.cn/disease/>) 花卉知识本体 Flower (www.flower-sh.cn/article_list.asp?c_id=74&page=4)，其相关信息如表 2 所示，对应资源是来自中国农科院作物品种资源研究所依据《中国粮食作物、经济作物、药用植物病虫害原色图鉴》、《中国农业百科全书》所制作的农作物病虫害知识以及来上海花卉网的资源，相关信息如表 3 所示。沿袭信息检索领域的传统，使用 Precision@n 和 Recall(n) 作为主要的评价方法和指标，其中 Precision@n 表示前 n 个结果中的查准率，Recall(n) 表示前 n 个结果中的查全率。计算如公式 (5) 和公式 (6) 所示。

$$precision @ n = \frac{\# of \text{ true results in top-}n \text{ results}}{n} \quad (5)$$

$$Recall @ (n) = \frac{\alpha \cap \beta}{\beta} \quad (6)$$

其中， α = 检索的前 n 个结果文档中实际相关的文档集合， β = 人工判断的检索结果中前 n 个结果。

表 2 实验本体统计信息

| Ont. Name | # of Concepts | # of Individual |
|-------------|---------------|-----------------|
| Flower | 1130 | 2,4000 |
| Soccer | 540 | 12300 |
| CropDisease | 2740 | 3,7300 |

表 3 文档集的统计数据

| 文档集 | 文档 总数 | 平均 | 实例平均标注 |
|----------|----------|-------|--------|
| | | 词数 | 文档数 |
| 花卉知识 | 1190 | ≈987 | ≈3.6 |
| 新浪国际足球新闻 | 4030 | ≈1135 | ≈5.4 |
| 农作物病虫害知识 | 1,1190 | ≈1129 | ≈4.3 |

同时，我们选择未经优化的原始概念检索方法作为基本对比方法 (RawMethod)。对考虑到大多数用户在检索过程中只考虑前 5-20 个结果，因此我们取 n=20 分别进行试验。在实验过程中，我们共构建了 100 个查询，对每个查询返回结果的前 20 个文档的相关性进行统计分析。

5.2 实验结果及分析

表 4 给出了两个方法在不同文档集上的实验结果评价。

表 4 3 个文档集上的实验结果

| Metric | Document Set | RowMethod | OurApproach |
|--------------|--------------|-----------|-------------|
| Precision@20 | 花卉知识 | 0.285 | 0.330 |
| | 新浪国际足球新闻 | 0.280 | 0.322 |
| | 农作物病虫害知识 | 0.283 | 0.338 |

| | | | |
|-----------|----------|-------|-------|
| | 花卉知识 | 0.202 | 0.254 |
| Recall@20 | 新浪国际足球新闻 | 0.185 | 0.240 |
| | 农作物病虫害知识 | 0.200 | 0.250 |

图1和图2给出了不同方法在3个文档集中的检索结果。

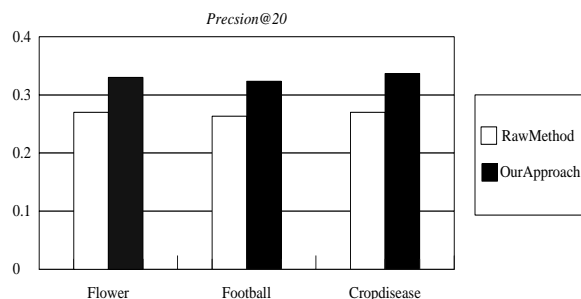


图1 两个方法在不同文档集上的Precision@20的比较

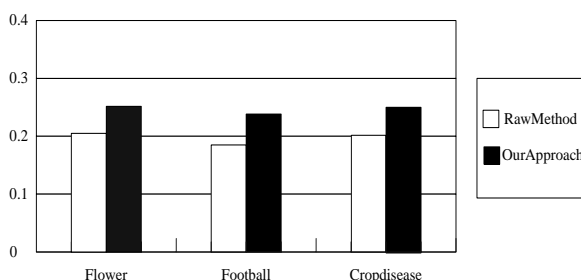


图2 两个方法在不同文档集Recall@20的比较

从图表结果来看，首先考虑问题规模对结果的影响，当本体的规模增大时，检索结果的查准率及查全率并没有太大的改变，由此可以看出本体规模对于检索效果影响不大。此外从文档集的数量来衡量，农作物病虫害知识的文档集最大，花卉知识的文档集最小，相差约10倍，但是检索效果相差小于2%，由此可以得出文档集的大小对于检索效果几乎没有影响。再对比原始概念检索方法，我们的方法在查准率及查全率上都有一定的提高，主要原因是方法考虑到本体中的上下文语义信息，而不是简单的统计本体知识标签出现概率。因此，方法对于查准率以及查全率的提高有一定的效果。

6 结论

随着语义网的出现，其在信息检索领域的应用-基于语义的检索相对传统的基于关键词检索是一个显著的进步，因为它能较好的解决基于关键词很难处理的词语问题。而其中词语-概念映射也是对于检索结果起着重要影响的一环。提出了一种词语-概念映射方法，该方法利用概念-文档与词语-文档两重关系，计算出词语与概念的相关度与置信度，再实现词语(集)映射到本体概念(集)。从实验结果来看，有效地提高了检索的效果。

参考文献

- [1] Hai Dong, Farookh Khadeer Hussain, Elizabeth Chang. A Survey in Semantic Search Technologies. 2008 Second IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2008): 403-408
- [2] 田萱,杜小勇,李海华.语义查询扩展中词语-概念相关度的计算.软件学报,2008年19卷8期: 2043-2053
- [3] www.w3.org/2001/sw/WebOnt/ WebOntology
- [4] Ciravegna F, Wilks Y. Designing adaptive information extraction for the Semantic Web in amilcare[G], Handschuh S, Staab S, eds. Annotation for the Semantic Web, Frontiers in Artificial Intelligence and Applications. Amsterdam: IOS Press, 2003: 112-127
- [5] Handsehuh S, Staab S, Maedche A. CREAM: Creating relational metadata with a component-based, ontology-driven annotation framework [G] Proc of the 1st Int'l Conf on Knowledge Capture. New York: ACM, 2001: 76-83
- [6] Alani H, Kim S, Millard D, et al. Automatic ontology-based knowledge extraction from Web documents[J]. Intelligent Systems, 2003, 18(1): 14-21
- [7] Lai Y, Wang R. Towards automatic knowledge acquisition from text based on ontology—centric knowledge representation and acquisition[C], Proc of the K—CAP 2003 Workshop on Knowledge Markup and Semantic Annotation(Semannot'2003).
- [8] Schutz A, Buitelaar P. RelExt: A tool for relation extraction from text in ontology extension[C], Proc of the 4th Int'l Semantic Web Conf(ISWC). Berlin: Springer, 2005; 593-606
- [9] <http://www.fao.org>
- [10] Vassilis Christophides, Grigoris Karvounarakis, Dimitris Plexousakis, et al: Optimizing taxonomic semantic web queries using labeling schemes[J]. Journal of Web Semantics 1, Issue 2, Elsevier (2003): 207-228
- [11] Tenier S, Toussaint Y, Napoli A, et al. Instantiation of relations for semantic annotation [C].Proc of Web Intelligence 2006. Los Alamitos: IEEE Computer Society, 2006: 463-472
- [12] Xu JX, Croft WB. Improving the effectiveness of information retrieval with local context analysis. ACM Trans. on Information Systems, 2000,18(1): 79 – 112.
- [13] 刘挺 车万翔 李生. 基于最大熵分类器的语义角色标注. 软件学报, 2007, 18 (03): 565-573
- [14] Yewang Chen, Wen Li, Xin Peng, Wenyun Zhao. An Improved Semantic Annotation Method for Documents Based on Ontology. CSWS(2009)
- [15] Furnas GW, Landauer TK, Gomez LM, Dumais ST. The vocabulary problem in Human-System communication. Communications of the ACM, 1987,30(11): 964 – 971.
- [16] Salton, G. and McGill, M. Introduction to Modern Information Retrieval.

- McGraw-Hill, New York (1983).
- [17] Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval. New York: Addison-Wesley-Longman, 1999.
- [18] Chang Y, Ounis I, Kim M. Query reformulation using automatically generated query concepts from a document space. *Information Processing and Management*, 2006,42: 453 - 468.
- [19] 张敏,宋睿华,马少平.基于语义关系查询扩展的文档重构.计算机学报,2004,27(10):1395-1401.
- [20] Gao JF, Zhou M, Nie JY, et al. Resolving query translation ambiguity using a decaying Co-Occurrence model and syntactic dependence relations. In: Järvelin K, Chairs P, Baeza-Yates R, Myaeng SH, eds. Proc. of the 25th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Tampere: ACM Press, 2002: 183 - 190.
- [21] Jang MG, Myaeng SH, Park SY. Using mutual information to resolve query translation ambiguities and query term weighting. Dale R, Church K, eds. Proc. of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics. College Park: Association for Computational Linguistics, 1999: 223 - 229.
- [22] 蔡怡峰, 彭鑫, 钱乐秋: 面向语义构件检索的交互式查询方案生成. 电子学报,2008,36(08):1631-1636

作者联系方式:

李文 Email:072021123@fudan.edu.cn
Tel: 13764155821 021-55075216
上海市杨浦区武川路 78 弄 29 号 101 室 200433