

文章编号:1001-9081(2010)06-1664-04

基于领域本体的政务信息检索系统

于 静¹, 吴国全², 卢 燮²

(1. 北京市地方税务局 第二稽查局,北京 100029; 2. 中国科学院 软件研究所,北京 100190)

(gqwu@otcaix.icas.ac.cn)

摘要:现有政务信息检索系统存在两个主要问题:一是采用基于关键词匹配的检索技术忽略了对用户检索条件的语义理解,缺乏对于文档实质内涵的准确描述;二是由于对政务信息领域知识的缺乏,用户不能很好地提出符合自己检索需求的检索条件。针对这些问题,提出了基于领域本体的政务信息检索方法,即通过引入本体,在文档和检索条件间建立一种基于本体的由本体中的词汇集组成的结构化的对应关系;设计并实现了相应的概念词抽取、检索条件扩展算法以及原型系统。实验结果表明,该方法在检索的查全率和查准率方面都有很大的提升。

关键词:本体;领域本体;政务信息;信息检索

中图分类号: TP311 **文献标志码:**A

Government information retrieval based on domain ontology

YU Jing¹, WU Guo-quan², LU Yi²

(1. The Second Audit Bureau, Beijing Local Taxation Bureau, Beijing 100029, China;

2. Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: There are two main problems in the existing government information retrieval systems. Firstly, the search technique based on keywords matching ignores the semantic information, which makes it unable to describe the document accurately. Secondly, due to the lack of domain knowledge, users are not clear about what they really want. To solve these problems, this paper presented domain ontology based government information retrieval approach, and designed the corresponding algorithm for concept extraction and query sentence extension. The experimental results show that the proposed approach improves the recall and precision ratio of the information retrieval.

Key words: ontology; domain ontology; government information; information retrieval

0 引言

政务信息检索系统作为整个政务信息公开平台中的一部分,具有不可替代的重要作用,它为民众方便、准确地定位自己所需的信息提供了保证。然而,现有政务信息检索系统存在两个主要的问题:一是基于关键词匹配的检索技术忽视了对于用户检索条件的语义理解,缺乏对于文档实质内涵的准确描述;二是由于对政务信息领域知识的缺乏,用户不能很好地提出符合自己检索需求的检索条件。这两个问题导致检索结果远远不能满足用户的要求。增加检索系统的语义检索能力^[1-2]是目前提升信息检索效果的一个有效途径。

本体^[3-4]作为一种能在语义和知识层次上描述信息系统的概念模型工具,通过概念间的关系来表达概念语义,其目标是捕获相关领域的知识,提供对该领域知识的共同理解,确定该领域内共同认可的词汇,并从不同层次的形式化模式上给出这些词汇和词汇间相互关系的明确定义,在基于知识的检索中有着广泛的应用。目前国内外已经有很多面向不同领域的基于本体进行语义信息检索的尝试^[5-6]。构建政务信息领域本体并将其用于政务信息检索系统中,具有重要的研究和现实意义。

本文提出了一种基于领域本体的政务信息检索方法。通过引入本体,在文档和检索条件间建立一种基于本体的由本

体中的词汇集组成的结构化的对应关系。关系的一边是将用户的检索条件基于本体扩展成结构化的检索条件,另一边是利用本体中词汇描述的结构化的文档标引。本体在政务信息检索系统中的应用主要体现在以下3个方面:

- 1) 检索条件的扩展和结构化检索条件的自动生成;
- 2) 利用本体中的概念,对政务信息进行标引,生成政务信息的实例库;
- 3) 通过对本体中定义的政务信息领域概念和关系的展示,对用户的检索进行提示和引导。

本文设计了相应的概念词抽取、检索条件扩展算法,实现了基于领域本体的政务信息检索系统,实验结果表明该方法在检索的查全率和查准率方面都有了很大的提高。

1 基于本体的政务信息检索

作为检索系统的核心,本体直接参与以下几个过程:概念映射、概念扩展、检索条件的生成和语义标注,图1给出了基于本体的政务信息检索步骤。其中,本体库的建立采用了基于政务主题词表的构建方法^[7],政务信息实例以结构化的形式存储于语义标注库中。

- 1) 概念词抽取。利用领域本体中的知识和一些简单的自然语言理解技术对用户提交的查询关键词或者自然语言表达的查询条件进行分析,提取主题词,并转换成由本体中概念

收稿日期:2009-11-26;修回日期:2010-01-12。

基金项目:国家科技支撑计划项目(2009BAH52B02);国家863计划项目(2007AA01Z149,2007AA04Z148)。

作者简介:于静(1977-),女,北京人,助理工程师,主要研究方向:电子政务、软件工程;吴国全(1979-),男,安徽合肥人,助理研究员,博士,主要研究方向:网络分布式计算、软件工程、面向方面的软件开发、业务流程管理;卢燚(1982-),男,山东兖州人,硕士研究生,主要研究方向:网络分布式计算、软件工程。

词所表达的一组检索概念。

2) 概念扩展。利用本体中的关系,对检索概念进行同义扩展、上下级扩展等一系列的扩展,获得一组更完备的检索概念。

3) 检索条件生成。将检索概念根据属性划分成不同的组,每一组对应结构化检索条件中的一项,根据对应关系,生成一组结构化的检索条件,并根据条件和检索意图的相关度进行排序。

4) 查询并返回结果。查询语义标注库并将结果返回用户,同时提示性地返回与用户检索条件中语义相关的本体概念实例及该实例与检索词的关系。用户可以根据这些反馈,调整检索条件,进而获得更确切的结果。

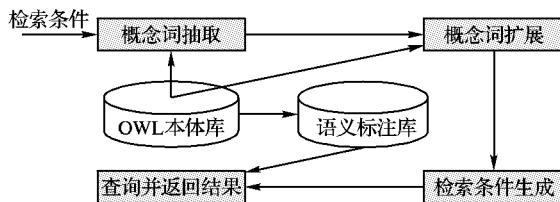


图1 基于本体的政务信息检索步骤

1.1 概念词抽取

在基于本体的信息检索系统中,为充分利用本体中的概念及关系,首要工作是抽取概念词,即将用户的检索条件和本体中的概念对应起来,将用户输入的检索条件转换成本体中的概念。概念词抽取可以将用户提交的由关键词或自然语言表达的检索条件映射到一个或一组概念词。

当用户录入查询条件时,可以键入单个关键词,如:查询表达式= \langle 出租汽车 \rangle ;也可以直接键入自然语言查询语句,如:查询表达式= \langle 出租汽车的相关收费情况 \rangle 。

对于单个关键词的检索条件,只需利用本体判断它是不是本体中的概念即可;对于自然语句查询,若用传统的分词方法进行分词,不能保证准确获取检索条件中蕴含的本体中的概念,因为本体中有些概念是分词后所得的基本词的组合。例如,本体中的概念“招生工作”就是由基本词“招生”和“工作”组合而成。这就要求分词后还要进行进一步组合等处理。

本文提出基于最大公共子串和经典分词的概念词抽取算法。最大公共子串匹配可以保证从自然语言查询语句中获取尽可能完备的概念,避免概念匹配不全的情况。但是,仅使用该方法可能会匹配不符合自然语言查询语句语义的概念词,本文基于分词法进行判断,过滤出不符合的概念词。算法描述如下:

输入:自然语言查询语句 QL 。

输出:本体概念词集 OCC 。

第1步 获取领域本体概念词汇表 $OC = (\omega_1, \omega_2, \dots, \omega_n)$ 和初始查询表达式 QL , 设 $OCC = \emptyset$ 。

第2步

```

for i = 1 to n {
    q = LCS(QL, ωi);
    if (q == ωi) {
        L1 = Split(QL);
        L2 = Split(ωi);
        if (L1 ∩ L2 != ∅)
            OCC = OCC ∪ {q};
    }
}
  
```

第3步 Output OCC 。

其中: $LCS(x, y)$ 输出 x 和 y 的最大公共子串, $Split(x)$ 输

出对 x 分词后的基本词组。该算法首先获取本体中所有概念的词汇表,对词汇表中的每个词,利用 LCS 算法同自然语言查询语句 QL 进行匹配,若匹配出的结果不是该本体中的概念词,则丢弃;若匹配出的结果是该本体中的概念词,则进行下一步的验证。验证的步骤是:分别对该匹配出的概念词进行分词,得到两组基本词,分别记作 L_1 和 L_2 ,若 L_1 和 L_2 没有交集,则说明获取的该概念词和自然语言查询语句没有任何基本概念的相关性,是一个不符合检索条件的概念词,需要将其丢弃。下面以两个具体的示例进行说明。

示例1 $QL =$ “招生的相关工作”。假设本体的概念中有“招生工作”这个概念词,记为 $\omega =$ “招生工作”。经过 $LCS(QL, \omega)$ 的运算,得到 $q =$ “招生工作”。经验证,该概念词符合条件,将其加入本体概念词集 OCC 。

示例2 $QL =$ “体育安全检查”。假设本体的概念中有“体检”这个概念词,记为 $\omega =$ “体检”。经过 $LCS(QL, \omega)$ 的运算,得到 $q =$ “体检”。显然该概念词和检索条件是没有语义关系的,应该丢弃。具体过程如下:对 QL 和 q 分别分词,得到两个基本词的集合,分别是 L_1 (体育,安全,检查) 和 L_2 (体检)。两个集合没有交集,说明通过最大公共子串匹配得出的概念词同检索条件分词后的各个基本概念没有任何公共部分,即不存在语义上的相关性。

1.2 检索条件扩展

检索条件扩展是指基于概念词,利用本体中定义的关系,找出能丰富和完善这些概念词的其他相关概念词,并分别用所有这些概念词作为政务信息实例各个属性的限定条件,将原有概念词所不能直接体现的条件信息补充进来,从而构成更精准的检索条件。

通过对用户的查询条件进行统计分析可得,对用户输入的查询进行概念词抽取后,可以抽取出一个概念词或者多个概念词。由于单个概念词和多个概念词在反映用户查询的语义方面存在差别,所以在进行查询扩展时要采取不同的策略。下面将分别对这两种情况下的扩展策略进行描述。

1.2.1 单个概念词的扩展

首先判断该概念词是属于哪个概念类别。

1) 若是一个类别概念的实例,记为 l ,则对其进行相关单位和相关主题词的扩展。扩展获得的概念词作为检索提示信息反馈给用户。

2) 若是一个单位的实例名称,记为 d ,则对其进行相关类别和相关主题词的扩展。扩展获得的概念词作为检索提示信息反馈给用户。

3) 若是一个主题词概念,则进行同义扩展、上下位概念扩展、相关概念扩展、相关类别扩展和相关单位扩展。对于扩展得到的概念词的处理策略是:同义扩展是最主要的,将扩展出的同义关系的概念词和原概念词放在一个集合,记为 ω^* 。

相关概念扩展、相关类别扩展和相关单位扩展获得的概念词,将不被扩展到检索条件中。因为对于单个概念词,其所反映的语义信息是有限的,除同义扩展外其他扩展获取的概念尽管同该检索词有一定关系,但若将其扩展后加入检索条件,将可能限制检索结果的完备性。因此,对于除同义扩展之外其他扩展获取的概念词,作为检索提示信息反馈给用户。

最终获得的概念词是 l 或 d 或 ω^* ,它们将被提交给检索条件生成部分作进一步的处理。

1.2.2 对于多个概念词的扩展

首先判断各个概念词分别所属的概念类别。若其中有类别概念实例或单位实例名称,分别记为 l_0 和 d_0 。下面是对其余的主题词概念进行扩展处理。

1) 进行相关类别和相关单位的扩展,获得排序后的相关类别和相关概念。然后进行同义扩展,获得各个概念词的同义词。上下位概念扩展、相关概念扩展获得的概念词,作为检索提示信息反馈给用户。

2) 进行相关类别和相关概念的扩展,获得多个相关类别和相关单位。按照它们与检索条件的相关程度进行排序,为下一步生成检索条件提供依据。

设 $W = (w_1, w_2, w_3, \dots, w_n)$ 代表由检索条件获取的多个主题词概念。首先定义一个相关度函数 $sim(d, W)$, 该函数用于计算由概念词扩展出的相关类别或相关单位 d 与 W 的相关度:

$$sim(d, W) = \left(\frac{1}{n} \sum_{i=1}^n length(d, w_i) \right)^{-1}$$

其中, n 为 W 中主题词的数目。

$$length(d, w_i) = \begin{cases} 0, & d \text{ 和 } w_i \text{ 间有相关关系} \\ 1, & d \text{ 和 } w_i \text{ 间没有相关关系} \end{cases}$$

下面是对相关类别的扩展:

对 W 基于本体获取 $w_i (i = 1, 2, \dots, n)$ 的一组相关类别, 记为 $L_i = (q_i^1, q_i^2, q_i^3, \dots, q_i^m)$ 。对 $L_i (i = 1, 2, \dots, n)$ 求并集得到 $L = \bigcup_{i=1}^n L_i$, 设其元素个数为 m 。然后对 L 中的元素 $l_i (i = 1, 2, \dots, m)$, 计算 l_i 与 W 的相关度, 记为 $s_i = sim(l_i, W)$ 。将 L 中的元素 l_i 相关度 s_i 由高到低进行排序, 元素排序后 L 记为 L' 。若检索条件中获取了 l_0 , 则将 l_0 加入 L' 且排放在第一位。图 2 描述了类别扩展的过程。

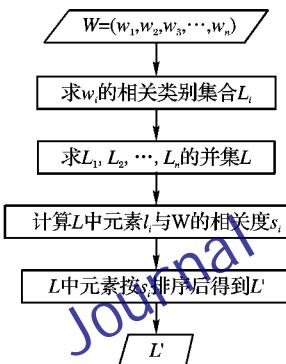


图 2 类别扩展流程

图 3 为一个具体的检索条件的类别扩展示例。

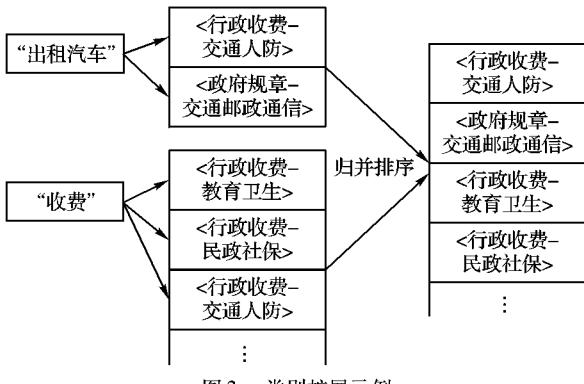


图 3 类别扩展示例

对相关单位的扩展采用相同的方法。记相关单位扩展后得到的排序的相关单位集合为 D' 。

下面对 W 进行同义扩展。对 W 中的每个概念词 $w_i (i = 1, 2, \dots, n)$ 进行同义扩展, 获得 $w_i^* = (w_i^1, w_i^2, w_i^3, \dots, w_i^k)$, 将 w_i 加入 w_i^* , 记为 w_i' 。用 w_i' 替换 W 中的 w_i , 得到同义扩展后的 W 记为 W' 。

至此, 得到 L' 、 D' 和 W' , 并作为下一步生成检索条件的基础。

1.3 检索条件生成

由查询条件扩展得到 L' 、 D' 和 W' 共 3 个集合分别对应本体中的类别、单位和主题词 3 个概念, 以对政务信息的所属类别、发布单位和关键词 3 个属性进行刻画。由于在组合条件录入模式中, 提供了除类别、关键词和发布单位以外的发布时间和文号两个检索条件, 最终生成的检索条件包括类别、单位、关键词、发布时间和文号。

检索条件的格式以及各项与本体中政务信息属性的对应关系如图 4 所示, 其中带阴影的条件表明不是由检索条件自动扩展的。

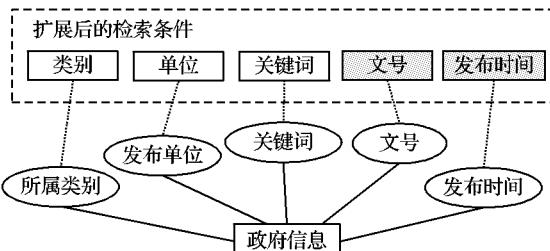


图 4 检索条件映射关系

鉴于政务信息的标引数据是以结构化的方式存储在数据库中, 检索方式采用基于数据库的查询, 因此需要将扩展后的检索条件实现为符合结构化查询语言的格式。生成结构化检索条件的过程如下:

1) 选定值 n 。选取 L' 中的前 n 个元素, 记为 $LS = (ls_1, ls_2, \dots, ls_n)$; 选取 D' 中的前 n 个元素, 记为 $DS = (ds_1, ds_2, \dots, ds_n)$ 。

2) 定义一个五元组, 记为 $s = (ls, ds, W', sj, wh)$, 其中 $ls \in DS, sj$ 和 wh 为直接获取的检索条件中的发布时间和文号值。

3) 将 LS 和 DS 中的元素依次填入 s 中的对应项, 获得一组五元组, 记 $S = (s_1, s_2, \dots, s_n)$ 。对 s_i 依 ls 和 ds 排序, 获得排序后的 S , 记为 S' 。

4) 生成查询语句。 S' 中的每个条件 $s_i = (ls, ds, W', sj, wh)$ 生成 SQL 语句的规则为: a) W' 中的 w_i^* 之间采用 or 连接, 对每个 w_i^* 中的 w_i^k 之间也采用 or 连接; b) ls, ds, W', sj 和 wh 之间采用 and 连接。

1.4 查询并返回结果

对 S' 中的五元组 s_i 生成的数据库查询语句, 执行数据库检索后返回一组结果, 记为 R_i 。由于 s_i 已经依据了扩展的精确程度进行了排序, 因此排在前面的结果集合更加符合用户的检索需求。在返回结果给用户时, 依据 S' 中的 s_i 顺序将 R_i 依次排列, 将 R_i 中的记录按发布时间排序, 发布时间晚的排前。

2 系统实现

图 5 给出了基于本体的政务信息检索总体架构, 主要包括本体构建、语义标注、查询扩展以及信息检索等 4 个功能模块。

本体构建 根据政务信息领域的特点, 使用基于政务主题词表的政务信息领域本体构建方法, 首先将发布单位、发布时间、所属类别、关键词和文号确定为政务信息的基本属性, 在添加单位、类别和主题词等概念和实例之后, 基于主题词表本身包含的各种关系进一步添加概念间的关系。本体开发工具采用 Protégé 2000^[8], 政务信息领域本体使用 Web 本体语言 OWL^[9] 进行描述并存储。

语义标注 采取人工标引的方式,向政务信息的标引数据库添加数据。其中,信息数据库中存储了政务信息的原始文档数据,包括政务信息的文件名、正文、摘要、类别、索引号、发布时间等信息。语义标注数据库中存储了对原始文档的标注信息,标引项包括政务信息的类别、主题词、发布单位和文号等。

查询扩展 对条件录入接口提供的用户初始检索条件进行一系列处理,包括概念词抽取、检索条件扩展后生成优化的、结构化的检索条件。使用 OWL 解析器——Jena^[10] 对 OWL 本体进行解析,获取检索扩展所需的本体信息。

信息检索 根据扩展后得到的结构化查询条件,在语义标注资源库中进行查询,得到查询结果并将其转换为实际文档数据返回给用户。

此外,该架构对外提供检索条件录入和政务信息录入两个接口。检索条件录入接口提供给用户一个检索条件的录入窗口;政务信息录入接口提供给政务信息的录入人员使用,所有信息的录入都将通过这个接口,政务信息的语义标注也通过该接口录入。

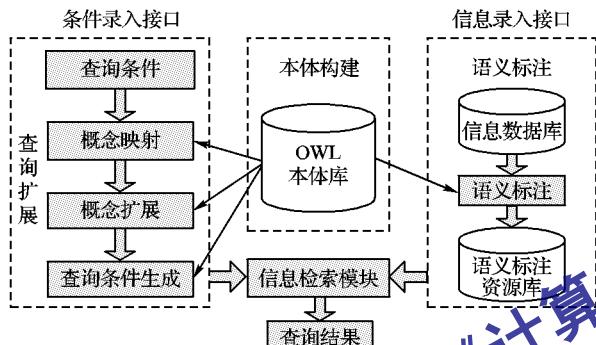


图5 基于本体的政务信息检索总体架构

3 实验分析与评价

实验使用参与开发的天津市政府信息公开平台(<http://www.tjzfxgk.gov.cn/tjep/>)作为检索系统,测试数据中共有测试文档 16 000 多篇,使用查全率和查准率^[11]作为主要评价指标,并与传统基于关键词匹配的检索方法进行了比较,实验结果如表 1 所示。

表 1 实验结果 %

检索方法	查准率	查全率
传统关键词匹配方法	52.3	63.3
基于领域本体的检索方法	78.8	82.3

下面通过具体的实例说明系统的检索效果。

实例 1 以<出租汽车>作为检索条件。该系统中关于“出租汽车”主题的文档 30 篇,原系统检索出 42 篇文档,其中 18 篇相关文档,本系统检索出 26 篇文档,其中 26 篇相关文档。

实例 2 以<交税>作为检索条件。该系统中包含相关主题的文档 6 篇,原系统检索出 0 篇文档,其中 0 篇相关文档,本系统检索出 5 篇文档,其中 5 篇相关文档。

实例 3 以<产业优化升级>作为检索条件。该系统包含相关主题的文档 3 篇,原系统检索出 0 篇文档,其中 0 篇相关文档,本系统检索出 2 篇文档,其中 2 篇相关文档。

实例 4 以<停车场的收费情况>作为检索条件。该系统包含相关主题的文档 10 篇,原系统检索出 664 篇文档,相关文档 10 篇。本系统检索出 22 篇文档,相关文档 7 篇。

首先,对检索条件中包含一个概念词的情况进行分析评

价。实例 1 表明检索查全率和查准率都有提升(查全率由 60% 提高到 87%,查准率由 42% 提高到 100%)。查全率提升是因为对检索条件进行了同义扩展,将“出租车”这个同义概念添加进来,丰富了检索条件;查准率提升是因为系统对文档做了标引,将仅仅含有“出租汽车”但并不反映租车主题的文档进行了排除。实例 2、3 之所以在原系统中检索不出结果,是因为所用的检索词是公文中一般不使用的词。本系统进行扩展后,分别将“纳税”和“产业升级”添加进检索条件,提升了查全率(实例 2 的查全率由 0% 提高到 83%,实例 3 的查全率由 0% 提高到 66%)。

其次,对检索条件中包含多个概念词的情况进行分析评价。实例 4 显示查准率提升(由 2% 提高到 32%),但查全率略有下降(由 100% 下降到 70%)。原系统检索出的是所有包含“停车场”和“收费”的文档,由于系统中包含“收费”关键词且与该检索条件不相关的文档数量较大,导致了较低的查准率。本系统依据提取出的“出租车”和“收费”两个概念词,扩展出相关类别“行政收费(交通人防)”和相关单位“**市物价局”,进而组合出更精确的检索条件,缩小了检索范围,过滤了无关文档,提高了系统的查准率。同时,过滤掉了 3 篇相关文档,使查全率略有降低。

最后,对检索提示功能进行评价。以<出租车>作为检索条件,系统在检索结果页面返回“公交车”、“机动车”、“客车”、“客运”和“公共交通”等相关概念词,返回“**市交通局”、“**市物价局”等相关单位实例,返回“行政收费—交通人防”和“政务规章制度—交通邮政通信”等相关类别。反馈信息对用户是一个提示和引导,用户可以根据这些信息调整自己的检索条件,让检索结果更加符合要求。

4 结语

本文针对政务信息领域的特点,分析了当前政务信息检索方法的缺陷,提出了基于领域本体的政务信息检索方法,基于本体中具有语义特征的规范词汇集,提出了对检索条件的概念词提取算法以及概念词扩展策略,从而支持检索条件到本体概念的全面映射。相比原基于全文匹配的系统,该方法在检索的查全率和查准率方面都有了显著的提升。

参考文献:

- [1] 宋炜,张铭.语义网简明教程[M].北京:高等教育出版社,2004:130-131.
- [2] 文坤梅,卢正鼎,孙小林,等.语义搜索研究综述[J].计算机科学,2008,35(5):1-4.
- [3] BERNERS-LEE T, HENDLER J, LASSILA O. The semantic Web [J]. Scientific American, 2001, 284(5): 34-43.
- [4] 邓志鸿,唐世渭.Ontology 研究综述[J].北京大学学报:自然科学版,2002,38(5):730-738.
- [5] FENSEL D. Ontologies: A silver bullet for knowledge management and electronic commerce [M]. Berlin: Springer, 2001.
- [6] 杨艳琴.领域本体查询体系结构和实现技术研究[D].哈尔滨:哈尔滨工业大学,2002.
- [7] 赵新力.综合电子政务主题词表[M].北京:科学技术文献出版社,2005.
- [8] Protégé[EB/OL].[2009-10-10].<http://protege.stanford.edu/>.
- [9] Web-Ontology (WebOnt) working group (Closed) [EB/OL].[2009-10-10].<http://www.w3.org/2001/sw/WebOnt/>.
- [10] Jena — A semantic Web framework for Java[EB/OL].[2009-10-10].<http://jena.sourceforge.net>.
- [11] BAEZA-YATES R, RIBEIRO-NETO B. Modern information retrieval [M]. New York: ACM Press, 1999.