

文章编号:1001-9081(2005)10-2434-02

## 基于语义网的电子政务文档智能检索

杨 芳, 杨振山

(同济大学 计算机科学与技术系, 上海 200092)

(yangfang@mail@126.com)

**摘要:**根据电子政务文档的特点,通过电子政务主题词表计算检索文档集和检索请求的特征值。讨论了检索文档集和检索请求的相似性计算,从而找到与检索请求匹配的文档。根据电子政务文档元数据的语义组织形式,研究电子政务文档元数据的检索问题。对所检索到的文档进行元数据语义组织,从而在语义推理的基础上实现智能检索。

**关键词:**电子政务文档; 检索; 语义网; 元数据

**中图分类号:** TP317.2    **文献标识码:**A

## E-Government document retrieval based on semantic Web

YANG Fang, YANG Zhen-shan

(Department of Computer Science and Technology, Tongji University, Shanghai 200092, China)

**Abstract:** According to the E-Government documents characters, the weight of the terms in documents and queries were calculated based on E-Government thesaurus. Similarity between query and documents was obtained through computing the similarity between weigh terms of documents and query, and most matching documents were provided. At the same time E-Government document metadata is organized in semantic web. E-Government document can also be retrieved as searching for metadata in semantic web. The documents retrieved is approached in semantic web with metadata to benefit logic reasoning.

**Key words:** E-Government document; retrieval; semantic web; metadata

## 0 引言

电子政务是信息化社会政府改革的必然选择,它使政府部门能够运用先进的计算机技术、通信技术和网络技术向全社会提供高效优质、透明和全方位的政府管理和服务。电子政务中的文档包含电子政务中所用到的各类文档信息,是电子政务的重要组成部分。如何使个体间或组织的部门间从电子政务丰富的文档资源中发现同手头工作相关的现有文档,分享实践经验,是电子政务文档智能检索的重要任务。

针对目前检索资源缺乏语义表示,而且没有进行语义检索,检索查准率低的问题,我们提出了基于语义网<sup>[1]</sup>的电子政务文档智能检索。该检索结合语义网的资源标注以及传统的概念检索技术,并根据电子政务领域中已有的分类体系和主题词表,在语义网的框架下进行知识推理和智能检索,提高检索的查准率。该智能检索实现框架如图 1 所示。

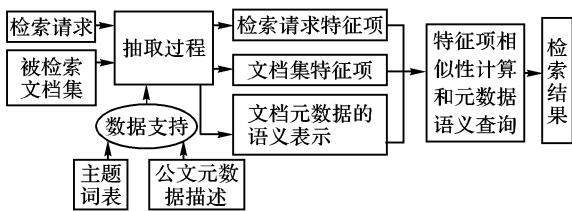


图 1 基于语义网技术的中文检索功能框架

## 1 数据支持

### 1.1 电子政务领域知识词表

领域知识词表代表领域知识的规范化语言,是领域内通

收稿日期:2005-04-30; 修订日期:2005-06-28

作者简介:杨芳(1976-),女,河南邓州人,博士研究生,主要研究方向:语义网在办公自动化中的应用; 杨振山(1935-),男,山东莱州人,教授,博士生导师,主要研究方向:办公自动化。

用的词表。

电子政务主题词表由范畴词、类别词、族首词和主题词多级层次组成。由于电子政务主题词表的规范化,我们这里使用数据库的形式进行存储。

### 1.2 文档集的语义元数据描述

对于电子政务中电子文档的检索来说,需要用到电子政务主题词表来标引和检索文档中的主题词; 电子政务文档元数据描述如文档的类型(函、报告、通报等)、文档的起草单位、发行时间、文档在网络中的位置等,从而有助于文档的检索要求,提高检索的效率。将电子政务文档中的元数据使用语义描述的形式表达,从而将文档以语义的形式组织起来。

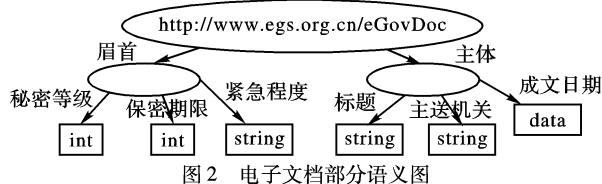


图 2 电子文档部分语义图

一般文献元数据的格式主要采用柏林元数据<sup>[2]</sup>,而对于电子政务文档来说,由于政务的特色,我们采用国家标准化组制定的元数据标准来描述元数据。然后使用语义网技术中的 RDF 资源描述框架(Resource Description Framework)<sup>[3]</sup>来描述电子政务文档的元数据以及元数据与元数据之间的关系。我们通过将文档的元数据资源转化到 RDF 的描述,从而达到语义之间的互操作。这是万维网所无法做到的,万维网只是提供了资源的显示形式,而没有给出资源的语义组织。

这里给出电子政务文档部分的语义描述图示,如图 2 所

示。在图 2 中, URL 地址为 < http://www. egs. org. cn/eGovDoc > 的文档,该资源其有两个属性:眉首和主体。对于眉首来说,其有三个性质:秘密等级、保密期限和紧急程度,它们的取值则由所对应的矩形所示;同样,对于主体来说,其有三个性质:标题、主送机关和成文日期,它们的取值范围由所对应的矩形所示。

## 2 抽取过程

被检索的文档集需要进行抽取检索词的预处理。抽取过程抽取出文档中的主题词,作为文档的标识。检索请求语句同样需要对检索语句进行抽取处理,以便得出检索请求的内容。而对于中文来说,由于词语之间没有间隔,所以需要首先进行分词,而且需要对所得到的词进行歧义检验,以便最大程度去除歧义。抽取过程具体如图 3 所示。

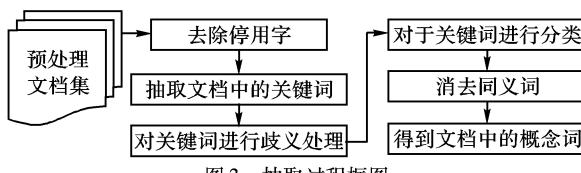


图 3 抽取过程框图

电子政务主题词表中不仅列出了电子政务领域的主题词,而且还给出了主题词间的关系。因此以电子政务主题词表为基础,经过以上的抽取过程,我们就可以得到检索文档集和检索请求中的主题词,并根据电子政务中主题词之间的关系,消去同义词,并将所得到的主题词以主题词表中的关系进行关联聚类。另外我们根据文档的固定格式抽取文档的元数据,然后将文档的元数据定义成 RDF 模型。

## 3 文档特征值计算

经过以上的抽取过程,我们就可以得到检索文档集和检索请求中的概念词。然后对得到的概念词进行特征值处理,得到文档中每一个关键词的权重,文档就可由这一系列检索词的权重值来表示,这些词就称为特征值。而查询请求经过同样的处理,也可以得到由一系列具有权重的检索词。通过对文档和检索请求权重词相似性计算,得到文档与检索请求之间的相似性。

特征值根据聚类原理进行计算<sup>[4]</sup>,这里使用了聚类相似性和非相似性两个度量值。

1) 聚类内部的相似性采用检索词在文档中的频率进行度量,使用的公式为:

$$tf_{i,j} = \frac{freq_{i,j}}{\max_l(freq_{l,j})} \quad (1)$$

其中  $freq_{i,j}$  表示概念词  $i$  在文档  $j$  中出现的次数,  $\max_l(freq_{l,j})$  表示文档  $j$  中最大频率的概念词  $l$  的频率值。 $tf_{i,j}$  值越大,表示概念词  $i$  出现在文档中的频率越高,越能代表该文档。

2) 聚类的非相似性是由整个文档集中出现索引词的文档篇数的倒数来计算,在文献[5]对此的计算有较多的研究。这里给出了一种被广泛使用并在测试中取得较好检索性能的计算方法。使用的公式为:

$$idf_i = \log_2 \frac{N}{n_1} \quad (2)$$

其中  $N$  表示整个文档集中文档的篇数,  $n_1$  表示出现索引概念词  $i$  文档的篇数。它反映了在其他文档出现较少的概念词越能代表该文档。

文档中概念词  $i$  在文档  $j$  中的权重值可由公式(3)计算:

$$w_{i,j} = tf_{i,j} \times idf_i \quad (3)$$

查询条件中索引检索词的权重<sup>[5]</sup>值可由出公式(4)计算:

$$w_{i,q} = (0.5 + 0.5 \times tf_{i,q}) \times idf_i \quad (4)$$

在公式(4)中, $tf_{i,q}$  表示检索词  $I$  在检索请求  $q$  中的相似度, $idf_i$  表示检索词  $i$  在文档集中与其他文档集的非相似度。

## 4 电子政务文档的检索实现

根据电子政务对文档的不同检索请求,我们将检索实现分为以下三类。

1) 当针对文档内容进行检索时,计算检索请求中检索词的特征值,然后与检索文档中的特征值进行相似性计算,得到检索结果。

对于所要检索的文档集来说,在检索之前需要对文档集进行预处理,即对文档集进行抽取处理,得到文档集中的概念词,并对每个概念词计算  $tf_{i,j}$ (公式 1) 和  $idf_i$ (公式 2),从而得到每个概念词  $i$  在文档  $j$  中的权重,得到文件  $j$  中每个概念词及其权重的表示(公式 3)。当文档集变化的时候,如增加文档时,则需重新计算  $idf_i$ ,从而重新计算每个文档中概念词的权重。这些都是预先进行处理的。对于具体的检索请求,抽取出检索请求中的检索词,然后计算其  $tf_{i,q}$ ,并根据文档集中检索词的  $idf_i$ ,来计算概念检索词  $i$  在检索请求  $q$  中的权重(公式 4)。

我们这里使用空间中的向量之间的夹角的余弦进行计算检索请求和文档集中的文档之间的相似度,计算公式如(5)所示。

$$\begin{aligned} sim(d_j, q) &= \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| * |\vec{q}|} \\ &= \frac{\sum w_{i,j} * w_{i,q}}{\sqrt{\sum w_{i,j}^2} * \sqrt{\sum w_{i,q}^2}} \end{aligned} \quad (5)$$

在公式(5)中,  $\vec{d}_j$  表示文档  $j$  中的特征值,  $\vec{q}$  表示检索请求的特征值。

2) 当针对文档元数据进行检索时,如检索文档的主送机关、发文日期、文档题目时,可以根据语义网中资源描述框架模型,针对检索请求关于检索文档的元数据内容,对语义网的资源进行查询,得到符合条件的内容。

由于我们这里对元数据使用 RDF 模型来描述,RDF 结构与元数据的结构相同。而针对文档元数据进行检索时,首先将查询元数据的请求转换成 RDF 结构。我们使用 RQL<sup>[6]</sup>查询语言来进行 RDF 结构元数据的查询,从而进行文档元数据的查询和推理。例如:

```

Select
  URL, Title
From
  {元数据} <Like> {条件}
  <source> {URL}
  <title> [Title]
...

```

上述 RQL 代码可以抽出元数据与条件相似的文档的 URL 地址和 Title 标题。

3) 当检索请求不仅对文档内容进行检索,而且还包括对  
(下转第 2438 页)

```

u32 count,loff_t *f_pos)
{ ..... /* 读函数则调用相应的 readb( ) 和 copy_to_user( ) 函数,
   与写函数同理 */
static int open_dpram( struct inode *inode,
   struct file *file )
{ .....                                /* 初始化 */
if ( ! request_mem_region(AT91_DPRAM,
   BUF_LEN * sizeof(u8), DEVICE_NAME) )
{ .....           /* 未申请到该内存空间时进行相应处理 */
//申请使用内存空间
base = ioremap(AT91_DPRAM, BUF_LEN * sizeof(u8));
//为设备内存区域分配虚拟地址
.....          /* 设置 DPRAM 读写时序 */
}
static int release_dpram( struct inode *inode, struct file *file)
{ ..... /* 释放相应资源 iounmap( ) 和 release_mem_region(); */

```

以上为 DPRAM 设备驱动的打开、读写、关闭函数的实现,然后通过以下标记化结构将其驱动的功能映射到前面的具体实现函数上:

```

static struct file_operations test_fops = {
  read: read_dpram,
  write: write_dpram,
  open: open_dpram,
  release: release_dpram
};

```

另外,在驱动程序初始化时必须通过 register\_chrdev( ) 注册。在加载该驱动前要使用 system("mknod /dev/设备名 c 主设备号 次设备号") 创建设备文件并为该设备分配设备号。

### 3 测试系统

本项目所搭建的测试系统包括 EPA 无线通信卡和 IO 模块控制卡两套 EPA 无线设备、一台 PC 机及一个灯箱,如图 4 所示。灯箱中的温度传感器与 AI 模块相连,将温度值传递给 AI 模块,并通过设备 A 发送到以太网上。设备 B 接收到此温度值后,将其与额定温度值相比较,如果低于额定温度值,则通过 AO 模块输出 4mA~20 mA 电流,控制灯箱内的可控硅模块,进而驱动灯箱内的灯泡,开始加热;如果高于额定温度值,则中断 AO 模块的输出电流,切断灯泡的电流输入,使灯箱内的温度下降,从而达到保持灯箱内温度恒定的目的。

(上接第 2435 页)

文档元数据的检索时,则我们首先将检索请求的文档特征值与检索文档中的特征值进行相似性计算;然后将满足相似性要求的文档进行元数据的 RDF 模型定义,作为 RQL 查询的基础;以检索请求中的元数据检索请求作为 RQL 的查询条件,对结果集的 RDF 模型进行查询,从而得到最后的检索结果。

由于我们使用元数据 RDF 模型,因此不仅可以得到满足检索要求的文档,而且还可以得到这些文档的元数据的 RDF 形式,使我们能够在语义网的基础上对这些文档元数据进行语义关系的推理,这些都是传统检索不能实现的特点。

### 5 结语

本文对电子政务文档和检索请求进行了主题词的抽取处理,得到文档的特征值,进行电子政务文档和检索请求中特征值相似性的计算。同时处理电子政务文档元数据以 RDF 模

实验证明,EPA 无线通信卡与 IO 模块控制卡之间数据传输稳定,这个系统运行效果良好,达到了预期目标,能够满足工业现场设备的通信要求。

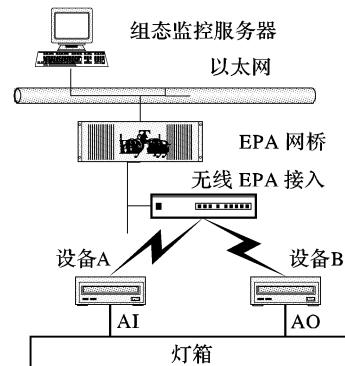


图 4 测试系统示意图

### 4 结语

EPA 系统是一种分布式系统,将分布在现场的若干个设备连接起来一起运作,共同完成工业生产过程和操作中的测量和控制。目前,无线局域网技术在工业控制中的应用已成为当今工业控制领域中的研究热点。但将无线技术应用于工业现场设备间的通信,并形成完整的分布式网络控制系统还是空白,也没相关的行业标准、国家标准和国际标准,专利也很少。因此,研究开发基于无线局域网的 EPA 通讯体系和工业现场控制设备原理样机及相关软件,形成基于 EPA 的分布式无线网络控制系统,具有很强的原创性。

#### 参考文献:

- [1] 高路,于海滨,王宏,等. EPA 网络体系结构[J]. 计算机工程, 2004, 17(30): 81~82.
- [2] 冯冬芹,金建祥,褚健. EPA 实时以太网标准化[J]. 工业以太网与现场总线, 2004, 8: 29~31.
- [3] 冯冬芹,袁剑荣,朱玲玲. 基于 EPA 的分布式控制系统网络通信模型[J]. 自动化仪表, 2003, 12(24): 60~63.
- [4] 朱斌,王平. 802.11b 接入 EPA 的安全策略[J]. 计算机应用, 2003, 12: 435~436.
- [5] RUBINI A. LINUX 设备驱动程序[M]. 第 2 版. 北京: 中国电力出版社, 2002.

型进行组织,使用 RQL 对文档元数据进行查询和推理,从而将查询的范围扩展到网络上,扩大检索文档的范围。

#### 参考文献:

- [1] BERNERS-LEE T, HENDLER J, LASSILA O. The semantic web, New York: Scientific American [J], 2001, 284(5): 34~43.
- [2] Dublin Core Metadata Element Set [DB/OL]. <http://dublincore.org/documents/dces/>.
- [3] MANOLA F, MILLER E. RDF Primer. W3C Working Draft [EB/OL]. <http://www.w3.org/TR/rdf-primer/>, 2002.
- [4] SALTON G. Introduction to Modern Information Retrieval[J]. Communications of ACM, 1983; 26(12), 1022~1036.
- [5] SALTON G, BUCKLY C. Term-weighting approaches in automatic retrieval. Information Processing and Management [J], 1988; 24(5): 513~523.
- [6] KARVOUNARAKIS G. The RDF Query Language (RQL) [EB/OL]. <http://139.91.183.30:9090/RDF/RQL>, 2003.