基于本体的 Web 页面聚类研究*)

谢红薇　颜小林　余雪丽
(太原理工大学计算机与软件学院　太原 030024)

摘　要　提出了一个基于本体的 Web 页面聚类系统原型,通过构建一个简单的搜索引擎并对结果进行聚类,大大节省用户发现所需信息的时间。同时将领域本体引入聚类系统中,提高了聚类效率和增强了聚类结果的可解释性。

关键词　聚类,本体,搜索引擎,向量空间模型,OWL

Research on Web Page Clustering Based on Ontology

XIE Hong-wei　YAN Xiao-lin　YU Xue-li
(College of Computer and Software Engineering, Taiyuan University of Technology, Taiyuan 030024, China)

Abstract　This paper brings out an ontology-based Web page clustering system, which greatly saves users time to find required information by building a simple search engine and also clusters its results. Introducing domain ontology to this clustering system improves the clustering efficiency and enhances the explainable of clustering results.

Keywords　Clustering, Ontology, Search engine, VSM, OWL

1　引言

随着互联网络的发展,Web 上电子文档的数目与日俱增,出现了"信息爆炸"的问题,互联网用户面对浩瀚的信息却难以找到所需的信息,因此如何从浩如烟海的 Web 资源中发现潜在的、有价值的知识已变得越来越重要。

Web 上的搜索引擎部分地解决了资源发现的问题,但由于它的搜索策略是基于关键字的匹配,缺乏对 Web 文档内容的全面把握和深层语义的正确理解,因此返回的结果远不能使用户满意。此外,搜索引擎一般将找到的相关万维网信息按照与查询的相关度从高到低排成一个线性列表,而且一个查询得到的搜索结果往往包括成千上万的万维网信息,所以用户得到的结果往往是个很长的线性列表。虽然搜索引擎已采用了各种方法来提高搜索精度,但搜索结果中仍然包括了大量与用户的查询无关的信息,从而使用户需要花费大量的时间去找到自己真正需要的信息,因此找到一种有效的方法解决这些问题具有重要意义。

这种重要性引起了许多研究者的注意,在这些方法中,主要分为两大类:一类是脱机的聚类,主要是通过网络蜘蛛(Spider)采用一定的爬行策略从各网站收集网页,然后再对其聚类,比较典型的有根据句法来给出一种相似度量[3,4];另一类是联机的方法,这种方法主要通过对查询结果聚类,主要是以主体为依据进行聚类。比较著名的 Vivisimo 就是采用的这种方法,还有相关的如 Grouper 与 Carrot 等。这些方法都在一定的程度上取得了好的效果。但在本质上没有对用户查询的意图很好理解,在聚类的时候决定聚类的个数时不能很好地确定,从而使得最后的聚类结果的可解释性方面比较差[5-7]。

因此本文提出一种基于本体的 Web 页面聚类系统,综合聚类方法和领域知识的优点,将用户提交的查询与领域本体匹配,提取背景知识来提供聚类的参数确定同时也增强了聚类结果的可解释性,通过将其应用到搜索引擎的背景下,大大方便了用户对所需信息的查找。

2　本体的应用

本体是共享概念模型的明确的形式化规范说明,它一方面可以帮助用户明确其信息需求,把未意识到的、未清晰表达的客观信息需求进一步显性化;同时让系统确定检索词在本体中的确定位置,从而帮助机器理解用户的检索意图,为用户提供更精确、更相关的知识和信息。

随着语义网技术的发展,出现了多种基于 Web 的本体描述语言,如 RDF(Resource Description Framework),DAML+OIL,OWL(Ontology Web Language,参考 http://www.w3.org/)等。OWL 作为 W3C 的推荐标准,是其所倡导的语义万维网(Semantic Web)的核心技术之一,意在提供一种语言,能够用于描述 Web 文档和应用中固有的类和类之间的关系。它通过定义类和类的属性来形式化一个领域,声明和定义对象和对象的属性,以及在 OWL 形式化语义允许的程度上对类(Class)和个体(Individual)进行推理。

3　基于本体的 Web 页面聚类的具体实现过程

3.1　基于本体的聚类挖掘框架

本文提出的基于本体的 Web 页面聚类系统,主要分为四个部分:页面采集与预处理,本体的应用,页面的特征表示,聚类。具体的系统模型结构如图 1 所示。

在模型中根据用户提交的查询返回一个结果集,同时解析用户提交的查询后将其匹配到一个领域本体,得到一个分类的背景知识,然后对结果集进行分类区分为 XML 或 HT-

*)基金项目:国家自然科学基金资助项目(60472093),山西省自然科学基金资助项目(20051035)。谢红薇　教授,博士研究生,主要研究方向为人工智能、并行计算;颜小林　硕士研究生,主要研究方向为信息检索、数据挖掘;余雪丽　教授,博士生导师,主要研究方向为智能软件工程、语义网。

ML，然后分别对页面进行解析，再根据领域本体对解析后产生的特征向量进行降维，从而为下一步的聚类做好准备，然后在分类背景知识的启发下对页面的特征表示向量进行聚类，并用用户提交的查询与领域本体匹配得到的信息作为聚类结果的表示依据，这样有效加快聚类的收敛速度，同时提高了结果的可解释性。

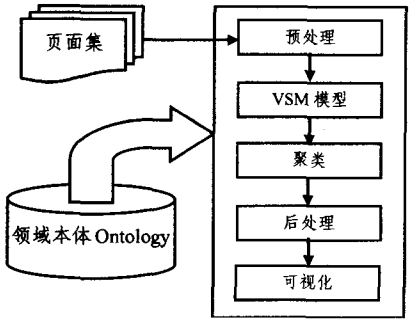

图 1　模型结构

3.2　页面的收集与预处理

页面的收集主要采用网络蜘蛛根据一定的策略从网站上采集相关的网页，然后对这些网页进行预处理。

Web 数据预处理就是去掉 Web 页面中与需要信息无关的其他内容，提取对分类和过滤有用的信息。主要分为两个过程：一个是对页面中控制符的分析，然后是怎样处理这些控制符。页面中控制符主要包括标题，即网页源代码中用〈‘T’ITLE〉和〈/TITLE〉标记的文字；关键字；页面描述，跟关键字类似的，在网页的头部说明中可以使用〈METHNAME–“DESCRIPTION”CONTENT–“……”〉的形式来描述页面内容；链接，链接元素用来描述两个文档或者文档和 URL 之间的关系；网页的正文部分：除了少数的专业网站外，大部分网站都是主要用自然语言书写。对于在 HTML 文档中出现的各种控制符号没必要把所有的都考虑进去，在实际训练的过程中，为了简化分析处理过程，仅考虑下面的控制符：TITLE（标题），META（置标），HREF（链接）等。

3.3　本体

3.3.1　本体的应用

在本文中挖掘时用到的本体侧重于一个较小的主题，对应于一个较小的领域应用。因此在此描述的是现实中的一个较小的领域应用，同时在应用中采用 OWL 来描述。下面是一个大学学校的 OWL 本体，其中的一个类层次如图 2 所示。

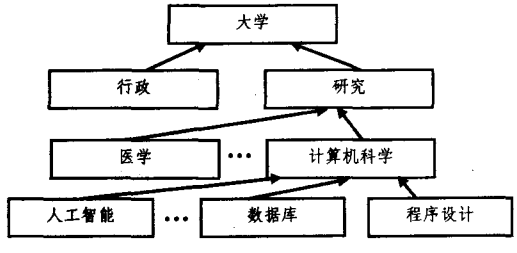

图 2　类层次

其中一个类的 OWL 语言描述如下：

〈owl:Class rdf:ID=“人工智能”〉

〈rdfs:subClassOf rdf:resource=“#计算机科学”/〉

〈/owl:Class〉

3.3.2　本体的匹配

将一个词汇的集合映射到一个领域本体时，希望得到一个公认的概念集，开始我们有一个集合 $T=\{t_1,t_2,\cdots,t_n\}$，通过将集合 T 与领域本体匹配后我们得到一个概念集 $C=\{c_1,c_2,\cdots,c_m\}$，这个领域本体用 OWL 来表示，先必须制定一些匹配的规则，根据用 OWL 表示的本体有三种对象，制定四种不同的映射，具体如下：

1）当集合 T 中 $t_i(i=1,2,\cdots,n)$ 与本体中类可以直接匹配时，将匹配到的概念 $c_j(j=1,2,\cdots,m)$ 直接加入集合 C 中。

2）当集合 T 中 $t_i(i=1,2,\cdots,n)$ 与本体中类属性可以直接匹配时，如果还和别的类匹配或那个类的实例匹配时，则遵循类的优先性最高，实例次之，属性最低，则将该类或实例所属的类作为匹配的概念输出，否则，则将该属性属于的类加入到集合 C 中。

3）当集合 T 中 $t_i(i=1,2,\cdots,n)$ 与本体中某个个体匹配时，当不与其它的类或属性匹配时，则直接将其所属的类加入集合 C 中，否则根据优先级将其类名作为概念加入集合 C 中。

4）当集合 T 中的元素与任何一个对象匹配不上时，则将其丢弃。

在语义相似度的计算中，在本文定义了一个新的计算概念之间的模型，具体的定义如下：一个概念 C_1 和一个概念 C_2，它们各自在本体中定义的属性为 $C_1=\{t_1,t_2,\cdots,t_n\}$；$C_2=\{t_1',t_2',\cdots,t_m'\}$，当 $\{t_1,t_2,\cdots,t_n\}\cap\{t_1',t_2',\cdots,t_m'\}=\{p_1,p_2,\cdots,p_k\}$，同时定义其中集合 $p=\{p_1,p_2,\cdots,p_k\}$ 中每个概念的权重分别为 $\{w_1,w_2,\cdots,w_k\}$，则可以得到两概念之间的相似性为：

$$\mathrm{Sim}(c_1,c_2)=(w_1+w_2+\cdots+w_k)/k$$

通过将两个概念的属性的集合相交，得到一个集合，然后根据每个集合中每个元素的权重相加后求平均值，这样把这个值作为两个概念之间的相似度。

3.4　基于本体的 VSM 表示模型

通过对文本集进行一定的特征提取后，得到每个文档的一个特征向量，得到特征向量后，通过将每个特征项与领域本体匹配，然后得到一个与本体匹配的新的特征向量。

通过分词后，每个文档表示成 $d_i(t_1,t_2,\cdots,t_n)$（其中 $i=1,2,\cdots,m$），将 $t_j(j=1,2,\cdots,n)$ 与领域本体匹配后得到一个新的向量 $Ont\_d_1=(c_1,c_2,\cdots,c_n)$。在传统的 VSM 模型中，文本空间被看作是一个有一组正交词条表示的的向量空间，每个文本表示为其中一个规范化特征向量 $V(d)=((t_1,w_1(d);t_2,w_2(d),\cdots,t_n,w_n(d))$，其中 t_1 为词条项，$w_1(d)$ 为 t_1 在 d 中的权重。TF-IDF 是一种常用的词条权重确定方法。由于 t_i 在文本中既可以重复出现又应该有先后次序关系，分析起来有一定难度，为了简化分析，可以暂不考虑 t_i 在文本中的先后次序并要求 t_i 互异（即没有重复）。这时可以把 $t_1,t_2,\cdots,t_n$ 看成一个 n 维的坐标系，而 $w_1,w_2,\cdots,w_n$ 为相应的坐标值，因此一个文本就表示为 n 维空间的一个向量，我们称 $V(d=(w_1,w_2,\cdots,w_n)$ 为文本 d 的向量表示或向量空间模型。其中每个词条的权重计算如下：

$$tfidf(d,t):=tf(d,t)\times\log\left(\frac{|D|}{df(t)}\right)$$

其中 D 为文档集，d 为任意文档，t 为一个文档中的词，$tf(d,t)$ 为词 t 在文档 d 中出现的频率，$|D|$ 为文档集的总数，$tf(t)$ 为词 t 在文档集中出现的次数，那么 $tfidf(d,t)$ 就为词 t 在文档 d 中的权重。

由于现在的词条都是通过把以前的词匹配到领域本体上得到的一个概念向量，因此会得到新的一个计算权重的公式：

$$cw_i = \sum_{k=1}^{n} ptw_k$$

其中 cw_i 为概念 c 在表示文档 d_i 时的权重，tw_k 为匹配前的词根据 TF-IDF 计算得到的权重，其中 p 为：

$$p = \begin{cases} 1 & tw_k \text{ 与概念 } cw_i \text{ 匹配} \\ 0 & tw_k \text{ 与概念 } cw_i \text{ 不匹配} \end{cases}$$

通过将所有的与概念匹配的词条的权重相加，那么得到一个新的表示一个文本的语义表示模型，每一个文本可以表示成 $V(d)=(cw_1, cw_2, \cdots, cw_n)$。

3.5 基于本体的 K-Means 聚类算法

在 K-Means 中需要初始的聚类参数，也就是确定聚类点的数目，这对用户来说是很难确定的，在这里通过将用户的查询匹配到领域本体上确定一个聚类点的数目 Ont_k，然后再聚类完成后结果的展示也以匹配的概念来表示，这样很好地解决了传统方法中聚类的结果可解释性不强的问题。算法的描述如下：

1)根据用户提交的查询的关键字匹配到一个领域本体上得到一个概念分类数 Ont_k，作为要生成的聚簇数目 k；

2)按某种原则选取 k 个初始聚簇中心，$C(c_1, c_2, \cdots, c_k)$，采用随机选取原则，设置初始迭代次数为 $r=1$；

3)对文本集合中没有分好类的每个文本 d_i，依次计算它与各个聚簇中心 c_j 的相似度 $\mathrm{sim}(d_i, c_j)$，这里将欧几里德距离作为相似度计算公式；

4)计算新的聚簇中心，新的聚簇中心为这一轮迭代中分到该聚簇中的所有文本特征向量的均值，即

$$c_j = \frac{1}{n_j} \sum_{d \in F_j} d$$

其中 F_i 为聚簇 c_j 的文本集合，n_j 为 F_j 中的文本数，d 为文本特征向量；

5)如果所有聚簇中心均达到稳定或者说准则函数收敛，结束；否则，$r=r+1$，goto(4)；

6)通过将中心点与概念类匹配，确定其对应的表示方式。

由此可知，该算法是基于迭代的过程。通常，初始点不同，聚类结果也不同。该算法运行速度快，时间复杂性为 $o(knr)$，其中 n 为总文本数，k 为聚簇数，r 为迭代次数。算法的缺点是必须事先确定 k 值，而在许多情况下，无法事先知道文本集合中的主题类别数目，在这里通过与领域本体匹配，确定了 k 的数目，大大地提高了收敛速度。

4 实验设计与实现

在研究当中采用网络蜘蛛从网站上采集页面，因为在研究当中主要采用的是一个大学本体作为领域知识，所以在采集的过程当中主要采集的页面从两个网站，中文网站采用的是浙江大学的站点(www. zju. edu. cn)，英文网站是采用的是 UIUC 的站点(www. uiuc. edu)，两个网站抓取的深度都设定 10 层，每层页面数为 200 个，通过将页面解析后构建一个概念的空间向量表示模型，然后再进行聚类，根据本体的概念的层次性的特点，我们采用改进的 K-Means 分块方法。

在系统的实现中使用 HP 工具 Jena 将用户的查询匹配到一个大学的一个领域本体，得到一个聚类的参数和最后结果聚类的一个概念表述方式；特征模型的构建方面我们使用概念构建的方法。系统的一个截图如图 3 所示。

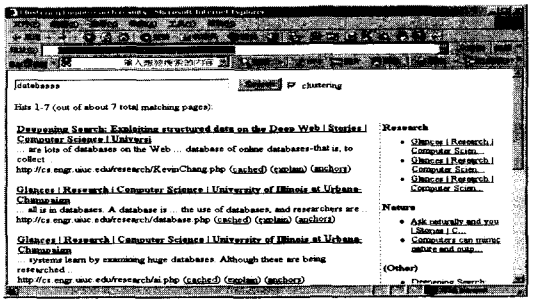

图 3 查询截图

结束语 本文通过将领域本体与文本的聚类方法结合，在搜索引擎返回的结果的背景下，对页面集进行聚类，根据用户提交的查询与领域本体的匹配，得到聚类参数和最后可解释的概念表述方式，在特征向量模型的构建方面通过将页面分词后的词语与本体匹配得到一个概念的向量模型，这样有效地提高了聚类的速度和增加了聚类结果的可解释性，并在一个引擎的环境下实现。

在研究中还存在一些问题。在中文的处理方面还有一定的局限性，现在没有一个很好能提供访问借口的中文词库，使得在匹配时有时不能很好地得到相应的聚类信息；在概念的匹配方面还需要找到更加有效的量化的方法。在后阶段，这些方面还有待进一步的研究。

参 考 文 献

[1] Tarau P, Mihalcea R, Figa E. Semantic Document Engineering with WordNet and PageRank // ACM Symposium on Applied Computing. 2005: 782-786

[2] Broder A, Charikar M A, Frieze, et al. Syntactic clustering of the Web, WWW6, 1997: 391-404

[3] Fang M, Garcia-Molina H, Motwani R, et al. Computing iceberg queries efficiently. VLDB, 1998

[4] Liu J, Wang W, Yang Jiong. A Framework for Ontology-Driven Subspace Clustering. KDD, Seattle, 2004: 623-628

[5] Menczer F. Combining Link and Content Analysis to Estimate Semantic Similarity. WWW2004, New York, 2004: 452-453

[6] Sara C, Jonathan M, Yaron K, et al. XSEarch: A Semantic Search Engine for XML // Proceedings of the 29th VLDB Conference. Berlin, Germany, 2003

[7] Beneventano D, Magnani S. A framework for the classification and reclassification of electronic catalogs // ACM Symposium on Applied Computing. 2004: 784-788

[8] Wang J, Chen Z, Li T, et al. Ranking User's Relevance to a Topic through Link Analysis on Web Logs // WIDM'02. McLean. 2002: 49-54