基于主题模型的结构化 Web 服务发现机制

陈江锋 于建军

(北京航空航天大学 软件开发环境国家重点实验室，北京 100191) (中国科学院 计算机网络信息中心，北京 100190)

摘 要：提出了基于主题模型(topic model)的结构化 Web 服务发现机制. 利用 LDA (Latent Dirichlet Allocation)生成概率模型，将 Web 服务(Web service)建模为结构化文本文档. 一个文档视作主题的概率分布，主题又由关键词的概率分布组成，从而提供基于主题的 Web 服务检索. 同时，利用 Web 服务的结构化特性，将 Web 服务描述文档表示为有向无环图，利用 n 阶谱核测量 Web 服务文档的相似度，实现 Web 服务结构化信息的发现. 通过实验分析对比，基于主题模型的结构化 Web 服务发现机制有效提高了 Web 服务发现的效率和精确率.

关 键 词：主题模型；Web 服务；n 阶谱核；服务发现

中图分类号：TP 391.4

文献标识码：A **文 章 编 号**：1001-5965(2008)06-0734-05

Topic model based structural Web services discovery

Chen Jiangfeng

(State Key Lab. Software Development Environment, Bejing University of Aeronautics and Astronautics, Beijing 100191, China)

Yu Jianjun

(Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: A structural Web services discovery approach based on a topic model was proposed. Each Web service would be modeled as a structural textual document using probabilistic model generated by latent dirichlet allocation (LDA). Every document could be seen as a multinomial random mixture of topics, and each topic had multinomial distribution over keywords, and thus a Web service retrieval based on the topic model was put forward. At the same time, the similar correlations in functionally structural Web services could be modeled as a directed acyclic graph (DAG) and be measured by using a gap-weighted n-spectrum kernel. Finally the experiment results show that several metrics for the classification and selection of services, such as success rate and executes efficiency, were improved via using our approach.

Key words: topic model; Web services; n-spectrum kernel; service discovery

随着 Web 技术的发展，基于面向服务的体系架构(SOA, Service Oriented Architecture)技术，利用 Web 服务封装现有网络资源，实现 Internet 上资源的集成和重用，目前已被工业界和研究机构广泛采用. Web 服务发现是实现 Web 服务应用的前提，只有发现满足用户需求的 Web 服务，才使得 Web 服务的组合、调用和监控成为可能. 因此，Web 服务发现是实现 Web 资源集成、共享的重要保障，Web 服务发现的结果将影响到上层应用系统业务逻辑的功能和质量，因此 Web 服务发现机制是 Web 服务生命周期的核心环节.

Web 服务发现本质上是针对 Internet 上海量 Web 服务，利用文本分类、模式匹配等信息检索技术，发现和获取满足用户需求的 Web 服务. 文本分类是在给定分类体系下，利用统计方法和

收稿日期：2007-03-04

基金项目：国家 973 基金资助项目(2005CB321901)

作者简介：陈江锋(1976—)，男，陕西韩城人，助理研究员，chenjf@buaa.edu.cn.

机器学习方法,根据文本内容确定文本类别的过程.应用文本分类技术的 Web 服务发现是用户通过输入关键词,通过 Web 服务搜索引擎提供的搜索服务,检索到具有相应关键词的 Web 服务候选集合的过程.这类方法能够较为准确地对新文档分类,但并不能很好地揭示关键词之间的关系,难以充分挖掘 Web 服务结构化信息和语义信息.潜在语义索引[1](LSI,Latent Semantic Indexing)以及 PLSI[2]等 bag-of-words 检索技术对文档和关键词间关系建模,速度快,且准确率较高,已广泛应用于自动文本分类和搜索引擎.但其只对文档和关键词间关系建模,并不能很好地体现关键词间的关系.主题模型建模关键词间关系,即实现主题识别.一些工作,如 LDA[3](Latent Dirichlet Allocation), Pachinko Allocation[4], Dynamic Topic Models[5]等认为文档由一些主题的分布组成,同时每个主题又是由关键词的分布来组成.这些工作一定程度上体现了文档—主题—关键词之间的关系,但其不能很好地建模 Web 服务的结构性.

针对 Web 服务的结构性,目前主要应用相似匹配机制进行建模.模式匹配技术,尤其是基于 XML 树匹配技术和核方法[6],提供了解决具备结构化信息的相似匹配通用解决方案.然而这些机制大多基于严格匹配机制,降低了检索的查全率,且没有很好地利用 Web 服务结构化特点,实际效果往往不佳.因此现有工作不能很好地建模 Web 服务的结构化特性和语义特性.

针对上述问题,本文提出了一种基于主题模型,结合结构信息的 Web 服务发现机制.Web 服务的发现过程分为两步:首先通过主题模型,结合 Web 服务特点和应用模式,分析 Web 服务与主题的组织关系,充分挖掘 Web 服务的潜在语义,实现 Web 服务的广泛获取,提高 Web 服务发现的查全率;其次,研究 Web 服务的结构特征,分析松散关联机制,引入核方法,提出结构化 Web 服务匹配模型,实现 Web 服务的精确检索.结合上述两种方法,提供基于主题模型的结构化 Web 服务发现机制,实现具有较高查询效率和精确度的 Web 服务发现.

1　基于主题模型的 Web 服务发现

与现有 Web 服务发现机制相同,本文首先将 Web 服务视为普通文本文档.由于 Web 服务采用 WSDL(Web Services Description Language)作为描述语言,提供了 Web 服务描述细节,因此本文采用 WSDL 作为 Web 服务的描述文本.

Web 服务是由一系列的服务接口组成,每个服务接口又由输入、输出参数和类型组成.因此,本文采用主题模型来建模 Web 服务的特性,充分挖掘 Web 服务的语义信息,提供 Web 服务的广泛检索机制.

主题模型的核心思想是认为一个文档是由一系列的主题分布组成的,而每个主题又是由一系列的关键词组成.区别于传统 bag-of-words 模型,主题模型强调文档是由文档—主题—关键词 3 层关系组成,而不仅仅是文档—关键词,即文档不是仅有单个主题组成,而是由多个主题组成.主题模型是一种生成概率模型,可以应用于文本数据、图像、生物图像以及其它多维数据的识别、分类和数据挖掘.

本文将采用目前被广泛使用的主题模型 LDA.LDA 将任意一个文档表示为一系列主题的组合,同时每个主题是由单词表的一系列关键词的分布组成.如图 1 所示,对于文档集合 D 中的任意一篇文档 w,LDA 使用如下方法生成:

1) N 符合泊松分布 $N\sim\text{Poisson}(\xi)$;

2) $\theta\sim\text{Dir}(\alpha)$;

3) 对于任一关键词 $w_n(1\leqslant n\leqslant N)$,选择一个主题 z_n 符合多项式分布,即 $z_n\sim\text{Multinomial}(\theta)$;在主题 z_n 条件下的多项式概率公式 $p(w_n\mid z_n,\beta)$ 下选择一个关键词 w_n.

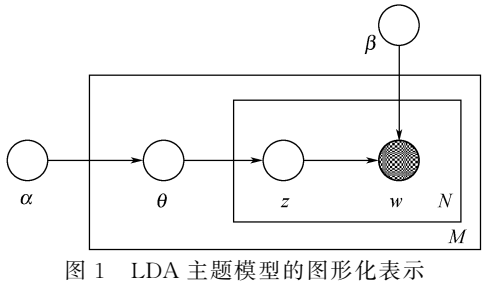


图 1　LDA 主题模型的图形化表示

给定参数 α 和 β,以及主题的混合参数 θ,文档中的主题 z 和关键词 w 的分布可用公式(1)来表示:

$$p(\theta,z,w\mid\alpha,\beta)=p(\theta\mid\alpha)\prod_{n=1}^{N}p(z_n\mid\theta)p(w_n\mid z_n,\beta)\qquad(1)$$

对于一篇文档,可用公式(2)来表示:

$$p(w\mid\alpha,\beta)=\int p(\theta\mid\alpha)\left(\prod_{n=1}^{N}\sum_{z_n}p(z_n\mid\theta)p(w_n\mid z_n,\beta)\right)\mathrm{d}\theta\qquad(2)$$

应用公式(1)和公式(2),可以利用 LDA 主

题模型,获取文档与主题,以及主题与关键词之间的关系.从图1的LDA模型图形化表示中可以看出,每个文档是由多个主题的分布形成,而每个主题又由多个关键词的分布组成,即LDA由3层结构组成:α和β是应用于文档集M的参数,以生成文档集;θ_d则用于生成文档集合中每个文档w_d;变量z_{dn}和w_{dn}用于生成每个文档w_d的每个单词.

根据LDA主题模型,可对现有Internet上Web服务文档进行训练和分类.首先,利用爬虫从Internet上获取Web服务描述文档,选择部分Web服务文档作为训练集,并对这些文档进行人工分类,分别利用LDA模型进行训练,使用测试集进行验证,获取最佳参数,以便对爬虫获取得到的新Web服务进行自动分类.

用户在查询Web服务时,输入关键词较少,基本在1~2个左右.使用传统的文本匹配技术,如空间向量模型,由于构造查询请求的关键词较少,不能很好地与Web服务文档进行匹配,不能有效地获取Web服务.基于主题模型的Web服务发现,可以将用户输入的关键词视作主题,利用主题模型,并根据主题与关键词之间的关系,映射为关键词空间向量,实现与Web服务文档的匹配,从而能够充分挖掘Web服务的潜在语义信息,广泛获取可能的Web服务.

2 基于结构信息的Web服务发现

主题模型揭示了文档—主题—关键词之间的关系,一定程度上挖掘了Web服务的潜在语义信息,提高了Web服务发现的查全率.然而,Web服务具有良好结构性,LDA主题模型并没有提供建模结构化特性的方法,因此为了充分利用Web服务结构化特性,本文将应用核方法来建模Web服务的相似性.

如第1节所述,Web服务可用WSDL文档来表示,而WSDL是由XML表示的.另一方面,Web服务间关系不似XML结构紧密联系,其组成关系是松散耦合,因此Web服务关键词之间的关系可用基于XML路径上的有向无环图来表示.为实现Web服务的分类,本文使用Web服务相似度来表示,在基于主题模型的Web服务发现的基础上,对获取得到的Web服务候选集,进行进一步的松散结构匹配,提高Web服务发现的精确度.下面详细介绍结构化Web服务发现机制.

定义1 假设XML树中的路径表示为$p=(n_1,n_2,n_3,\cdots,n_k)$,存在指标$i=(i_1,\cdots,i_u)$满足使得对于$j=1,\cdots,|u|$,有$u_j=p_{i_j}$,其中$1\leqslant i_1\leqslant\cdots\leqslant i_{|u|}\leqslant k$,则称由$i$指示组成的路径上的有序结点集合为$p$的路径序列,用$n$-ps表示.

路径序列在p的位置由i给出.用$|i|=|u|$表示指标个数,则n-ps的长度$i_{|u|}-i_1+1$,即路径序列所包含的路径中的结点个数.

路径序列为原有路径中结点集合的子集,保持了路径的结构信息,当$|i|=k$时,则路径序列为原有路径.同时也可以通过路径序列集合还原为原有Web服务描述文档.例如〈BuyTicket,TravelTicketDetail,TicketInfo,Date〉是TravelTicketService中的一条路径,(BuyTicket,TracelTicketDetail,Data)则是这条路径上的3-ps,即长度为3的路径序列.

定义2 给定两棵XML树D_1和D_2,使得对于任何一个结点$x\in N(D_1)$,最多只有一个结点$y\in N(D_2)$与之对齐.对齐结点集合表示为$N_A\subseteq N(D_1)\otimes N(D_2)$,对于$(x,y)\in N_A$,且$x\in N(D_1)$,$y\in N(D_2)$,称$x$和$y$为$N_A$中的对齐结点,表示为$x\leftrightarrow y$.其中$\otimes$表示笛卡尔积.

假设$N(D_1)=\{n_1,n_2,n_3\}$,$N(D_2)=\{n'_1,n'_2,n'_4\}$,如果有$n_1\leftrightarrow n'_1$,$n_2\leftrightarrow n'_2$,则称$\{(n_1,n'_1),(n_2,n'_2)\}$为有效的对齐结点集合.

按照上述定义,将存在大量的对齐结点.如D_1的结点数目$N(D_1)=n$,D_2的结点数目$N(D_2)=m$,则存在$n\times m$个可选择的对齐结点.本文定义了一个非负函数Score(x,y)来计算候选的对齐结点之间的相似程度,并通过求解$\max\left(\sum^{\min\{m,n\}}\text{Score}(x,y)\right)$值,得到一种能最大程度反映结点间相似关系的对齐方式.求解$\max\left(\sum^{\min\{m,n\}}\text{Score}(x,y)\right)$的实质是在二分图中寻找最大匹配问题,本文通过经典的Ford-Fulkerson算法来求解;首先将结点的元素属性值当作一个包含多个关键字的文本文档,然后通过结点间的文本相似度来计算Score(x,y).

定义3 给定两棵XML树D_1和D_2,以及对齐结点集合N_A,对于D_1中一个路径序列$u=(u_1,\cdots,u_n)$,若存在D_2中的一个路径序列$v=(v_1,\cdots,v_n)$,对于$1\leqslant i\leqslant n$,满足$u_i\leftrightarrow v_i$,则称u和v为等价路径序列,记为$u\Leftrightarrow v$.

路径序列中的结点按照文档树中结点的顺序关系来排列,从而反映了Web服务描述元素或属性之间的层次关系.本文基于以下假设:两个

Web 服务越相似,则它们将拥有越多等价路径序列.

基于结构化核计算等价路径序列集相似度,首先要将等价路径序列集中的数据项嵌入到向量空间中,本文称之为嵌入空间. 嵌入空间 F 坐标的指标定义在路径序列上,给出了一个维数为 n 的向量空间,用 ϕ 表示特征映射:ϕ: ps $\mapsto$ $(\phi_u(\text{ps}))\in F$. 可以有不同的映射 ϕ 来供选择,本文通过 n 阶谱核来定义映射 ϕ. n 阶谱核原来是用于比较 2 个字符串有多少个公共长度为 n 的子序列. 本文通过泛化 n 阶谱核,使之适用于更为普通的结构化对象——XML 文档树上. 即把一棵 XML 文档树当作一个字符串,将路径序列当作串的子序列. 这样,求解 XML 文档树的相似度问题可以转化为求其路径序列集的相似度.

考虑下面情况,假设有路径序列(A,B,C)存在于路径$\langle A,B,C,D\rangle$,$\langle A,B,D,C,E\rangle$和$\langle A,F,G,H,B,C\rangle$中,本文认为第 1 个路径更重要,因为从中提取的路径序列是连续的. 为了处理非连续的路径序列,本文引入一个加权衰减因子 $\lambda\in(0,1)$来表示路径序列中的间隙,而不是对所有路径序列分配相同的权重. 路径序列用指标 $i=(i_1,\cdots,i_u)$来唯一标识;用 $l(i)$表示该路径序列中第一个结点到最后一个结点的在相应路径中的距离. 考虑间隙加权的 n 阶谱核中,本文用 $\lambda^{l(i)+1}$表示该路径序列的出现加权. 例如,从$\langle A,F,G,H,B,C\rangle$提取路径序列$(A,B,C)$,其权重可表示为 λ^6.

定义 4　间隙加权的 n 阶谱核. 输入空间为 $I=\sum^n$(路径序列集合),特征空间为 F,间隙加权的 n 阶谱核的映射 $\phi^n:D\mapsto(\phi^n_{\text{ps}}(d))_{\text{ps}\in I}\in F$ 可表示为 $\phi^n_{\text{ps}}(D)=\sum_{i:\text{ps}=D(i)}\lambda^{l(i)+1}$,$\text{ps}\in\sum^n$. 其中,$i$ 是指标,ps 是路径序列,λ 是衰减因子. 相对应的核函数定义为

$$\kappa_n(s,t)=\frac{\langle\phi^n(s),\phi^n(t)\rangle}{\|\phi^n(s)\|\cdot\|\phi^n(t)\|}=\frac{\sum_{\text{ps}_s\Leftrightarrow\text{ps}_t}\phi^n_{\text{ps}_s}(s)\phi^n_{\text{ps}_t}(t)}{\|\phi^n(s)\|\cdot\|\phi^n(t)\|}$$

其中,$\text{ps}_s\in\Psi^n(s)$;$\text{ps}_t\in\Psi^n(t)$.

该函数定义了 2 棵文档树在 n 长度路径序列上的相似程度. 显然,有 $\kappa_n(s,t)\in[0,1]$. 如果 2 棵文档树结构严格一致,则它们的相似度为 1. 如果结构完全不一样,即不存在等价序列,则相似度为 0. 显然,有 2 棵文档树结构越相似,两者的等价路径序列越多,$\kappa_n(s,t)$值则越大.

2 棵文档的结构相似度可以定义为一系列间隙加权的 n 阶谱核的核函数之和,即 Web 服务的结构相似度表示为

$$\text{Sim}(S_a,S_b)=\sum_{i=2}^{k}\alpha_i\times\bar{\kappa}_i(D_a,D_b)\qquad(3)$$

其中,$k\geqslant 2$;D_a,D_b 分别表示 Web 服务 S_a 和 S_b 所表示的文档树;α_i 表示各个 n 阶谱核的核函数的权重,有 $\sum_{i=2}^{k}\alpha_i=1$.

得到 Web 服务间相似度 $\text{Sim}(S_a,S_b)$后,则利用 KNN(K-Nearest Neighbour)文本分类算法对 Web 服务进行分类,从而获取不同功能的 Web 服务集合. 用户在查询时,转化为分类问题,选出最佳分类集合,挑选合适的 Web 服务候选集,实现 Web 服务发现.

3　实验分析和评测

本文从 Internet 上获取2 460个有效的 Web 服务,提取 WSDL 文档,并对这些 WSDL 文档中的元素和属性中所表示的短语进行过滤和分析,获取得到548 142个关键词(包含1 536个单词). 在实验过程中,本文使用1 600个 WSDL 文档作为训练集,剩余的 WSDL 文档作为测试集.

基于主题模型的 Web 服务发现机制,由于采用 LDA 模型,需要训练 LDA 的相关参数. 本文从训练 Web 服务集合中,设置 20～48 个分类,分别进行参数学习,获取最优参数. 整个实验过程在一台 2.8 G XEON CPU,4 GB 内存的 PC 服务器运行,其中训练集上的参数学习平均时间大约为 30 min,测试集则平均约为 16 min. 表 1 给出了 5 个示例主题,并且给出分布其上的 TOP 10 关键词.

表 1　5 个样例主题的关键词分布

样例主题	关键词									
主题 1	zipcode	address	forecast	city	country	location	hazard	state	weather	locate
主题 2	message	short	mobile	send	content	receive	gateway	number	call	user
主题 3	isbn	author	price	publisher	query	pay	account	buy	bank	credit
主题 4	Travel	service	Trip	Airline	Flight	weather	Room	Hotel	Schedule	Time
主题 5	soap	parameter	structure	response	firmware	deadline	sale	request	in/out	service

表1可见，主题1～主题4，能够很好的体现当前Web服务的语义信息，即它们分别提供了天气服务、短消息服务、售书服务以及旅游服务；而在主题5中，则很难从中获得当前主题的内容. 这是因为，WSDL定义了很多关键词，如“service”，“soap”，“in/out”. 在实际Web服务发现中，需要将WSDL关键词过滤.

表2和表3分别给出了数量最多的6个Web服务分类，以及应用主题模型LDA和结构信息n阶谱核的Web服务发现查全率/查准率.

表2 基于主题模型的Web服务发现的查全率/查准率

服务类别	服务数量	查全率/%	查准率/%
weather	73	82.67	72.33
zipcode	57	87.39	73.20
stock	39	80.54	70.67
book	21	80.87	69.86
math	18	76.32	69.55
sms	12	89.33	78.43

表3 基于结构的Web服务发现的查全率/查准率

服务类别	服务数量	查全率/%	查准率/%
weather	73	79.74	86.15
zipcode	57	81.56	80.37
stock	39	73.21	81.09
book	21	78.25	86.40
math	18	74.96	79.26
sms	12	79.69	87.67

从表2和表3可以看出，LDA模型相对n阶谱核，具有较差的查准率，这是因为主题模型忽略了Web服务的结构特性. LDA模型将根据主题与关键词的关系，将文本相似、但结构不相似的Web服务选定作为候选集. 而n阶谱核则具有较差的查全率，这是因为n阶谱核关注于结构信息，部分文本的语义信息被忽略了. 因此，为实现精确的Web服务发现，本文将采用基于主题模型的结构化Web服务发现机制，即采用主题模型，进行广泛的Web服务获取，并在其基础上，应用n阶谱核进行Web服务的精确匹配.

4 结束语

针对Internet上大规模Web服务发现问题，本文通过主题模型，挖掘Web服务的潜在语义信息，实现Web服务的广泛获取；通过n阶谱核，建模Web服务的结构化信息，实现Web服务的精确匹配. 同时本文通过实验设计和分析，证明了基于主题模型的结构化Web服务发现机制具有较高的查询效率和精确度.

参考文献（References）

[1] Deerwester S, Dumais S, Furnas G. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 1990, 41:391－407

[2] Hofman T. Probabilistic latent semantic indexing[C]//Gey Fredric, Hearst Marti, Tong Richard. Proceedings of the 22nd Annual International ACM SIGIR Conference. New York: ACM, 1999:50－57

[3] Blei D M, Ng A Y, Jordan M. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3(1):993－1022

[4] Li W, McCallum A. Pachinko allocation: DAG-structured mixture models of topic correlations[C]//Cohen William, Moore Andrew. Proceedings of the 23rd International Conference on Machine Learning. New York: ACM, 2006:577－584

[5] Blei D M, Lafferty J D. Dynamic topic models[C]//Cohen William, Moore Andrew. Proceedings of the 23rd International Conference on Machine Learning. New York: ACM, 2006:113－120

[6] Shawe-Taylor J, Cristianini N. Kernel methods for pattern analysis[M]. United Kingdom: Cambridge University Press, 2004:344－372

（责任编辑：文丽芳）