中 文 信 息 学 报
第 13 卷 第 3 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 13 No. 3

一种基于奇异值分解的双语信息过滤算法*

路海明 徐晋晖* 卢增祥 李衍达

清华大学自动化系 北京 100084 *清华大学计算机系 北京 100084

摘要 本文提出了一种基于 SVD(奇异值分解)[1]的双语信息过滤[2]算法,将双语文档进行了统一的表示,使得适应于单语过滤的算法可以方便地用于双语过滤,同时对文档向量进行了压缩,滤去了噪声。在应用方面,将双语过滤算法用于互联网上的个性化主动信息过滤。

关键词 双语信息过滤 SVD 互联网 Bookmark 服务

一、引言

互联网上的信息中英文并存,国内用户,需要同时获取中英文信息,而现在的个性化信息服务只是针对一种语言进行的。用户建立了中文下的用户模型之后,希望获得自己喜好的英文资源。即能够同时为用户提供两种语言的信息过滤。双语过滤指"根据用户在一种语言里的兴趣表达,产生两种语言的推荐结果",如给定表达用户需求的中文向量,能够向用户推荐满足用户需求的中英文两种语言的文档。为解决这个问题,主要有基于机器翻译的算法和基于统计学的算法。

1.1 基于机器翻译的算法[2]

1. 翻译关键词

用户的需求用中文关键词查询向量表示,将每个关键词翻译成英文,形成英文向量,再去查询英文文档,进而返回推荐的英文文档,实现双语过滤。

2. 翻译文档

用户的需求用中文关键词查询向量表示,将所有英文文档进行全文翻译,产生中文文档,用中文关键词向量查询翻译产生的中文文档,产生用户需要的中文文档,对应的英文文档推荐给用户,实现双语过滤。

机器翻译方法的主要优点是通俗易懂、实现方便、效率较高,但机器翻译本身仍存在很多困难,导致信息过滤的结果也不理想。当前信息过滤的算法本身也有较大误差,人们似乎还能够容忍机器翻译带来的误差,基于机器翻译的信息过滤还占有一定的市场。

1.2 基于统计学的算法[4]

统计学算法,采用训练文档集,其中的每篇中文文档都有对应的英文文档。其匹配的基本方式仍旧是根据矢量空间模型 Vector Space Model(VSM)[5],用户需求和文档都表示成向量,利用余弦计算相似度。

例如用户的需求向量

* 本文于 1998 年 12 月 11 日收到

$$\vec{q} = (q_1, q_2, \cdots, q_n)^t$$

表示文档的向量

$$\vec{d} = (d_1, d_2, \cdots, d_n)^t$$

则两者之间的相似度

$$sim(\vec{q}, \vec{d}) = \cos(\vec{q}, \vec{d}) = \frac{\sum_{i=1}^{n} q_i d_i}{\sqrt{\sum_{i=1}^{n} q_i^2 \sum_{i=1}^{n} d_i^2}}$$

基于统计学的双语过滤方法有 Peseudo - Relevance Feedback(PRF)[6]和 Generalized Vector Space Model: GVSM[7],本文根据 Latent Semantic Indexing: LSI[8][9]的思想,提出基于 SVD(奇异值分解)的双语统一向量表示方法,并做了论述。

首先说明 PRF 方法、GVSM 方法和 SVD 方法

1. PRF 方法

根据用户的中文需求向量,在训练集的中文文档中查询出最相似的 k 篇中文文档,它们在训练集中有相应的 k 篇英文文档,根据这 k 篇英文文档,获取用户的英文需求向量,用该向量去获取满足用户需要的英文文档,从而实现双语过滤。

2. GVSM 方法

训练集中的所有中文文档,形成 term-document 矩阵 A_{mxn}(通过矩阵 A_{mxn}表示 n 篇文档 document,共有 m 个词 term 用于表示这 n 篇文档,每篇文档用一个 m 维列向量表示,列向量的每个元素分别表示 m 个词的权重)。所有英文文档,形成 term-document 矩阵 B_{exn},表示有 e 个词,n 篇文档。

如果按照列观察矩阵 A_{mxn},可以将矩阵 A_{mxn}看作由 n 个向量组成,每个向量表示一篇文档,即通过不同词的出现情况表示文档。另一方面,如果按照行观察矩阵 A_{mxn},可以将矩阵 A_{mxn}看作由 m 个向量组成,每个向量表示一个词,即通过该词在不同文档中出现的情况表示一个词。PRF 方法采用的是第一种观察矩阵 A 的方法,将词作为基本元素。Wong 提出 GVSM 方法,采用的是第二种观察矩阵 A 的方法,将文档作为基本元素。

例如,对于中文查询向量 $\vec{q}$,通过变换 $\vec{q}' = A^t\vec{q}$, 形成 n 维向量 $\vec{q}'$,即用训练集文档的不同权重表示中文查询向量 $\vec{q}$,对于英文文档向量 $\vec{d}$,变换为 n 维向量 $\vec{d}' = B^t\vec{d}$,也是用训练集文档的不同权重表示英文文档向量 $\vec{d}$。则相似度 $sim(\vec{q}, \vec{d}) = \cos(A^t\vec{q}, B^t\vec{d})$,这样便实现了双语过滤。

3. SVD 方法

令 A 是一个 m * n 维的实数矩阵,则存在 m * m 和 n * n 的正交阵 U_{mxm}和 V_{nxn},使得 $A = U\sum V^T$, $\sum$是 m * n 的对角阵,其主对角线上的元素$\sigma_{11} \geq \sigma_{22} \geq \cdots\cdots \geq \sigma_{hh} \geq 0$; h = min(m, n),见[1]。

σ_{kk}称为矩阵 A 的奇异值,U 和 V 分别叫做矩阵 A 的左奇异阵和右奇异阵。奇异值σ_{kk}包含了有关矩阵 A 秩的特性。

设矩阵 A 为一个 term-doc 矩阵。通过奇异值分解可以有效地将矩阵 A 进行压缩,并保持各文档的相似度基本不变。

两篇文档的相似度可以用代表这两篇文档的向量夹角的余弦值来描述,对于列归一化(指任意一列的向量的模为 1)的矩阵 A 来说,文档的相似度可以由 A^TA 来表示,这样 $S = A^TA$ 中的任何一个元素 s_{ij}都表示文档 i 和文档 j 的相似性。

由奇异值分解得 $A = U \sum V^T$,令 $C = \sum V^T$,显然满足 $C^T C = A^T A$,由于 $\sum$的元素满足$\sigma_{11} \geq \sigma_{22} \geq \cdots\cdots \geq \sigma_{hh} \geq 0$; $h = \min(m, n)$,即越靠后的奇异值对于整体的影响越小,可以考虑只取矩阵 $\sum$的前 $k(k <= h)$个奇异值构成矩阵 $\sum_1$来代替矩阵 $\sum$, $\sum_1$为 $k * n$ 的矩阵,满足

$(\sigma_{11}^2 + \sigma_{22}^2 + \cdots\cdots + \sigma_{kk}^2) / (\sigma_{11}^2 + \sigma_{22}^2 + \cdots\cdots + \sigma_{hh}^2)$接近 1,见[1]。

这样,令 $D = \sum_1 V^T$代替矩阵 A 将可以很好地保持文档之间的相似度,同时实现了对 term - doc 矩阵 A 的压缩,产生的矩阵 D 为 $k * n$ 维,这样一篇文档本来用 m 维向量表示,压缩为用 k 维向量表示,经验表明, m 通常大于 10 000, k 通常小于 1000,因此压缩率显著。

在原始的 term-doc 矩阵 A 中,代表每篇文档的 term 向量由所有关键词组成,其中一些关键词的出现对于评价文档并没有本质作用,它们的存在实际上起到了噪声的作用。通过奇异值分解,只抽取起主导作用的成分,从而起到噪声过滤的作用。

4. 基于 SVD 的统一向量方法

本文将提出一种基于奇异值分解的双语过滤方法,将双语文档变换为统一的向量,从而使得传统的单语过滤算法可以应用在双语过滤上,避免了由于机器翻译造成的误差,同时通过奇异值分解,降低了文档向量的维数,提取了文档的主要内容,滤去了噪声,也提高了过滤速度。

5. 几种统计学算法的比较

基于 SVD 的统一向量方法,缺点是需要进行 SVD 运算,有比较大的计算量,优点是在以后的计算中,由于向量进行了大幅度压缩,计算量减少,同时滤去了噪声。

PRF 和 GVSM 方法,其优点是无需进行 SVD 运算;其缺点是在以后的运算中,由于向量未进行压缩,计算量较大,同时有噪声的影响。

二、基于 SVD 的双语过滤

2.1 双语矩阵 W 的产生

记

$$W = \begin{bmatrix} A \\ B \end{bmatrix}$$

其中,矩阵 A 为中文的 term-doc 矩阵,其每一列表示一篇文档;矩阵 B 为英文的 term-doc 矩阵。

文章要被表示成向量,首先需要进行预处理,对待英文,预处理的步骤为:去掉 stopword,如 the、that 等,而后进行 stemming,如将 played, playing 变为 play;对中文要增加切词的工作。预处理之后,文章变为一个词集(terms)。词集中的每个词(term)都需要一个权值,通常是采用词(term)的 TFIDF(Term-Frequency Inverse-Document-Frequency)[5]加以计算的。在一个给定的文章集中,使用 TFIDF 方法,文章 i 中词 k 的权值

$$dw_{ik} = tf_{ik} * [\log_2(n) - \log_2(df_k + 1)]$$

其中 tf_{ik}为词 k 在文章 i 中的频率, df_k 为包含有词 k 的文章数, n 为总文章数。

矩阵 W 中的 A 和 B 都是根据 TFIDF 产生的列归一化矩阵。

2.2 矩阵 A 与矩阵 B 的关系

定理 1:

$$A^T A \approx B^T B$$

证明:由于矩阵 A 的第 i 列表示第 i 篇文档,因此矩阵 A 的第 i 列和第 j 列的内积表示中

文文档 i 和中文文档 j 的相似度,即矩阵 A^TA 中元素(i,j)的值。同理,矩阵 B 的第 i 列和第 j 列的内积表示英文文档 i 和英文文档 j 的相似度,即矩阵 B^TB 中元素(i,j)的值。而英文文档 i 跟中文文档 i、英文文档 j 跟中文文档 j 在本质内容上是同一篇文档,只是用不同的语言表达,实际实验也表明其相似度基本一致,即 $A^TA \approx B^TB$

证毕!

为了理论分析的方便性,假设 $A^TA = B^TB$

定理 2:若 $A^TA = B^TB$,且从 A 到 B 的变换为线性变换,B = CA,则 C 为正交矩阵。

证明:$A^TA = B^TB = (CA)^T(CA) = A^TC^TCA$,有 $C^TC = I$,所以 C 为广义正交阵,即矩阵 A 可以通过左乘一个广义正交阵 C 变换为矩阵 B。

证毕!

2.3 利用矩阵 W 的奇异值分解,实现双语统一压缩向量

定理 3:设 $A = U\sum V^T$, $W = EFG^T$, B = CA, C 为广义正交阵,取 H 为 E 的前 m 列,有

$$H^T\begin{bmatrix} A \\ 0 \end{bmatrix} = H^T\begin{bmatrix} 0 \\ B \end{bmatrix}$$

证明:由前面 2.1 中及 B = CA, C 为广义正交阵,有

$$W = \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} A \\ CA \end{bmatrix}$$

$\therefore W^TW = A^TA + A^TC^TCA = 2A^TA$

又 $A = U\sum V^T$, $W = EFG^T$

$\therefore$

$$F = \sqrt{2}\begin{bmatrix} \sum \\ 0 \end{bmatrix}$$

$$G = V$$

设

$$E = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix}$$

则由 $A = U\sum V^T$, $W = EFG^T$, G = V 及

$$w = \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} A \\ CA \end{bmatrix}$$ 有:

$$\begin{bmatrix} U\sum V^T \\ CU\sum V^T \end{bmatrix} = \begin{bmatrix} U_{11}\ U_{12} \\ U_{21}\ U_{22} \end{bmatrix}\sqrt{2}\begin{bmatrix} \sum \\ 0 \end{bmatrix} V^T$$

$\therefore, U_{11} = \sqrt{2}U/2 \quad U_{21} = \sqrt{2}CU/2$,

因此,

$$E^TW = \begin{bmatrix} \sqrt{2}U^TA \\ 0 \end{bmatrix} \quad E^T\begin{bmatrix} A \\ 0 \end{bmatrix} = \begin{bmatrix} \sqrt{2}U^TA/2 \\ U_{12}^TA \end{bmatrix} \quad E^T\begin{bmatrix} 0 \\ B \end{bmatrix} = E^T\begin{bmatrix} 0 \\ CA \end{bmatrix} = \begin{bmatrix} \sqrt{2}U^TA/2 \\ U_{22}^TCA \end{bmatrix}$$

$\therefore$当只取矩阵 E 的前 m 列时,即取 H 为 E 的前 m 列,有

$$H^T\begin{bmatrix} A \\ 0 \end{bmatrix} = H^T\begin{bmatrix} 0 \\ B \end{bmatrix}$$

证毕!

根据该定理,对于到来的文档,无论是中文还是英文,通过左乘 H^T 将得到统一的文档表示向量。例如,如果是同一篇文档的中英文表示,则得到的向量将基本一样。

推论 1:如果矩阵 W 的 m 个奇异值中(最多为 m 个,其真实取值为矩阵 A 的秩和矩阵 B 的秩中较小的值,称为 m 是为了表达和理解方便),前 k 个奇异值占的比重较大,当取矩阵 Q

为矩阵 E 的前 k 列时,将有

$$Q^T\begin{bmatrix} A \\ 0 \end{bmatrix} \approx Q^T\begin{bmatrix} 0 \\ B \end{bmatrix}$$

由于通常 $k \ll m$,例如当 m 为 10000 左右时,k 为 200 左右,这样一方面将中英文进行了统一的表示,另一方面实现了数据压缩。该定理可以通过 1.2 和定理 3 得到。

通过前面分析,左乘矩阵 Q^T 可以将矩阵 A 和 B 统一起来并且压缩,即任意给出一个单语的向量 Y,无论其是哪种语言,只要左乘矩阵 Q^T 就会得到一个统一的 k 维压缩向量 T,这样就可以采用各种单语过滤的方法进行双语过滤了。

2.4 统一向量的误差分析

前面假设 $B = CA$,且 C 为正交阵,这样就会保证 $A^TA = B^TB$,事实上,对于给定的双语文档和相应的关键词,并不能保证矩阵 A 和矩阵 B 的协方差的完全一致性,只能保证其协方差的大致一致性,即 $A^TA \approx B^TB$。现在的问题是在协方差不一致的情况下,是否还可以通过左乘矩阵 H 使得仍能保证 $H^T\begin{bmatrix} A \\ 0 \end{bmatrix} \approx H^T\begin{bmatrix} 0 \\ B \end{bmatrix}$

不妨设矩阵 W 的上半部分不是 A,而是 $A+\alpha$,当 α 跟 A 比较可以忽略时,W 的奇异值分解可以近似为跟忽略 α 的奇异值分解的结果比较接近,这样

$$E^T\begin{bmatrix} A+\alpha \\ 0 \end{bmatrix} = \begin{bmatrix} \sqrt{2}\,U^T(A+\alpha)/2 \\ U_{12}^T(A+\alpha) \end{bmatrix} = \begin{bmatrix} \sqrt{2}\,U^TA/2 \\ U_{12}^TA \end{bmatrix} + \begin{bmatrix} \sqrt{2}\,U^T\alpha/2 \\ U_{12}^T\alpha \end{bmatrix} = Y_1$$

$$E^T\begin{bmatrix} 0 \\ B \end{bmatrix} = E^T\begin{bmatrix} 0 \\ CA \end{bmatrix} = \begin{bmatrix} \sqrt{2}\,U^TA/2 \\ U_{22}^TCA \end{bmatrix} = X_1$$

因此误差就表现在 $\sqrt{2}\,U\alpha/2$ 上,由于我们只关注矩阵 Y_1 与矩阵 X_1 的前半部分,即

$$Y_2 = \sqrt{2}\,U^TA/2 + \sqrt{2}\,U^T\alpha/2$$

$$X_2 = \sqrt{2}\,U^TA/2$$

若没有 $\sqrt{2}\,U^T\alpha/2$,则矩阵 Y_2 跟矩阵 X_2 保持了完全一致性,现在有了 $\sqrt{2}\,U^T\alpha/2$ 使得矩阵 Y_2 跟 X_2 有了差别,只要 X 不是很大,则 $\sqrt{2}\,U^T\alpha/2$ 将也不会很大,这样仍能近似地保持 Y_2 和 X_2 的一致性。

三、双语过滤算法在 Bookmark 服务中的应用

3.1 Bookmark 服务简介

传统浏览器上的 Bookmark 功能存在一些不足,由于浏览器上的 Bookmark 功能是针对某个具体的浏览器设计的,用户在一个浏览器上建立了自己的 Bookmark,但是当他更换浏览器(如从 Netscape 换为 IE)或更换机器时,原来建立的 Bookmark 可能不再存在,这样会给用户带来许多不便。使用我们提供的信息推荐服务,我们设置在用户端的 agent 将能使用户把 Bookmark 存储在设立的网络服务器上,无论用户在全球任何地方、用任何机器、采用任何操作系统、使用任何网络浏览器,只要能连接到我们的服务器上,就都拥有自己统一的 Bookmark,这样将大大方便用户,因为用户不必为失去自己的 Bookmark 而烦恼。

用户 Bookmark 信息反映了用户的信息偏好,用户 Bookmark 中的文档为用户所喜欢的,我们根据用户的信息偏好对用户进行信息推荐,用户对推荐的内容进行评价,评价信息反馈给我们,以便进一步优化用户模型。利用 Bookmark 进行用户需求的获取和信息推荐符合用户

的使用习惯,很容易被接受。现在,我们的 Bookmark 服务已经开通,可以访问如下地址:http://166.111.72.50,免费使用我们的服务。

3.2 双语过滤算法的应用

从前面的分析可知,我们需要根据双语矩阵 W,产生矩阵 E,进而产生矩阵 Q,对于新的的文档向量(无论中英文),通过左乘 Q^T 就可以得到统一的压缩向量。这里面存在一个问题,那就是新的文档向量不一定属于双语矩阵 W 中的向量,而奇异值分解产生的矩阵 H 是否还适应?如果矩阵 W 足够大,并且具有广泛的代表性,则通过左乘 Q^T 得到的压缩向量误差较小。因此初始的双语矩阵 W 的选择对于系统性能的影响较大。

在 Bookmark 服务中,用户 Bookmark 中的文档为用户所喜欢的,我们向用户推荐的文档中有的用户喜欢,有的用户不喜欢,这就形成了正负两方面的训练集,我们将这些文档分别产生压缩向量,形成压缩向量下的正负两方面的训练集,根据 SVM[10,11]方法,产生分类线。我们的 Robot 启动搜索引擎主动获取网页,对于获取的网页,无论中英文,通过左乘 H^T 得到统一的压缩向量,将该压缩向量跟 SVM 的分类线向量比较,判断用户是否喜欢,如果认为用户喜欢,则推荐给用户。

3.3 实验数据和实验结果

在 3.2 中提到,初始的双语矩阵 W 的选择对于系统性能的影响较大,如果矩阵 W 足够大,并且具有广泛的代表性,那么在很大范围内其过滤效果也会较好。但是,形成足够大的双语矩阵需要很多双语文档和计算机运行资源。因此,目前只在小范围内进行了实验。

1. 原始文档收集

原始文档分为两类:一类是关于数据挖掘(Data Mining)的,另一类是关于多媒体广播(Multimedia Broadcast)的。两类文档各选取了 30 篇,其中 20 篇用来作为训练集,10 篇用来作为测试集。这些文档分别有英文文档和相应的中文文档,即共有 120 篇文档。

2. 关键词选取

关键词的选取采用计算机整理与人工选取相结合的方法,这样一方面可以提高效率,另一方面可以保证选取的关键词有相当的代表性和分类效果。抽取的 172 个关键词如下:

抽取的英文关键词(90 个):

algorithm, artificial, Bayesian, capture, character, classification, cluster, databases, decision, discover, domain, empirical, expert, extraction, feature, gain, induction, intelligent, KDD, knowledge, learn, machine, mine, model, network, neural, pattern, recognition, reduction, representation, retrieval, rule, structure, study, template, text, ATM, Audio, broadcast, camera, CD, channel, communication, compression, conversion, DAB, data, decode, demodulation, delivery, digital, disseminate, DMB, DVB, DVD, encode, frame, Frequency, GSM, image, information, interactive, ISDN, Kbps, medium, modulation, MPEG, multicast, multipoint, multimedia, PSTN, radio, rate, sample, satellite, signal, simultaneous, stereo, sound, synchronization, telecommunication, telephone, television, terminals, transfer, TV, UHF, uncoded, VHF, video

抽取的中文关键词(82 个):

算法,人工,贝叶斯,获取,特征,分类,聚类,数据库,决策,发现,领域,经验,专家,提取,归纳,智能,KDD,知识,学习,机器,挖掘,模型,网络,神经,类型,识别,压缩,再现,规则,结构,学习,模板,文本,ATM,音频,广播,照相机,CD,通道,通信,转化,DAB,数据,解码,解调,发

送,数字,散布,DMB,DVB,DVD,编码,祯,频率,GSM,图像,信息,交互,ISDN,Kbps,媒体,调制,MPEG,多点,多媒体,PSTN,声音,速率,采样,卫星,信号,同步,立体声,电信,电话,电视,终端,传输,TV,UHF,VHF,视频

3. 形成 W 矩阵

由程序自动完成 60 ×2 篇文档的关键词抽取,按照 TFIDF 的算法形成各文档的关键词向量,并进行归一化,即向量的模为 1,形成矩阵 $W_{172\times60}$。

4. 通过奇异值分解进行压缩

通过奇异值分解对 W 矩阵进行压缩,取前 21 个奇异值。这样,原始的 60 ×2 篇文档压缩为向量长度为 21 的 60 ×2 个向量。

说明:按照前面的算法,应当由一个足够大的双语矩阵进行奇异值分解,然后跟这 60 篇长度为 90 和 60 篇长度为 82 的文档运算,形成压缩向量。但是,由于双语文档有限,只好将这 60 ×2 篇文档向量形成的矩阵直接压缩,以提高精度。

5. 实验结果

这 60 ×2 个长度为 21 的向量包括 30 ×2 篇关于数据挖掘的文档和 30 ×2 篇关于多媒体广播的文档。用其中的 20 ×2 篇关于数据挖掘的文档和 20 ×2 篇关于多媒体广播的文档作为训练集,用其中的 10 ×2 篇关于数据挖掘的文档和 10 ×2 篇关于多媒体广播的文档作为测试集。采用 SVM 方法进行分类,分类结果如下:

10 篇关于数据挖掘的中文文档: 9 篇分在数据挖掘中, 1 篇分在多媒体广播中

10 篇关于数据挖掘的英文文档: 10 篇分在数据挖掘中, 0 篇分在多媒体广播中

10 篇关于多媒体广播的中文文档: 2 篇分在数据挖掘中, 8 篇分在多媒体广播中

10 篇关于多媒体广播的英文文档: 1 篇分在数据挖掘中, 9 篇分在多媒体广播中

对于数据挖掘类文档,其指标如下:

查准率(Precision) = (10 + 9)/ (10 + 9 + 2 + 1) = 0.86

查全率(Recall) = (10 + 9)/ (10 + 10) = 0.95

对于多媒体广播类文档,其指标如下:

查准率(Precision) = (8 + 9)/ (8 + 9 + 1 + 0) = 0.94

查全率(Recall) = (8 + 9)/ (10 + 10) = 0.85

6. 计划进行的实验

下一步计划使用 Reuters - 21578 数据集*。有许多实验利用 Reuters - 21578 数据集进行,其实验结果容易跟别人比较。其中主要利用的切分方式之一为:"ModApte"切分,利用其中 9603 篇作为训练集,3299 篇作为测试集。在这种切分下,选择文章最多的 10 类进行实验。去掉 stopword 和经过 stemming 后,共有 10083 个英文 terms。主要工作要将这些文档翻译成中文,需要比较好的机器翻译工具。当然,只要找到足够多的双语文档,就可以做进一步的实验。同时,将该结果与 PRF 和 GVSM 方法进行比较。

四、结论

* 数据集由 David Lewis 编辑,它来源于 1987 年路透社新闻。
http://www.research.att.com/~lewis/reuters21578.html

通过 SVD 进行双语过滤,压缩了数据,滤去了噪声,提高了速度,方便了用户,避免了由于机器翻译造成的误差。通过 Bookmark 服务获取用户信息需求和进行双语主动信息推荐,符合用户的使用习惯。

参 考 文 献

[1] 张贤达. 现代信号处理. 北京:清华大学出版社,1995,68 - 69

[2] Carbonell Jaime G, Yang Yiming, Frederking Robert E *et al*. Translingual Information Retrival: A Comparative Evalution. IJCAI97. 1997,708 - 714

[3] Belkin N J, Croft W B. Information filtering and information retrieval: Two sides of the same coin. Communication of ACM 35,1992,12(Dec.): 29 - 38

[4] Vapnik V. The Nature of Statistical Learning Theory. New York: Springer, 1995

[5] Salton G. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Pennsylvania: Addison-Wesley, 1989

[6] Buckley C, Salton G, Allan J *et al*. Automatic Query Expandsion Using SAMRT: TREC 3. In: Overview of the Third Text Retrieval Conference (TREC - 3), 1995,69 - 80

[7] Wong S K M, Ziarko W, Wong P C N. Generalized Vector Space Model In Information Retrieval. In: ACM SIGIR Conference on Research and Development in Information Retrieval. (SIGIR '85), 1985,18 - 25

[8] Deerwester S, Dumais S T, Furnas G W *et al*. Indexing by Latent Semantic Analysis. In: J Amer Soc Inf Sci 1, 1990,6:391 - 407

[9] Dumais S, Landauer T, Littman M. Automatic Cross - Linguistic Information Retrieval using Latent Semantic Indexing. In: Proceedings of SIGIR - 96, Zurich, Auguest 1996

[10] Joachims T. Text categorization with support vector machine. Technical Report. LS VIII Number 23, University of Dortmund, 1997

[11] Cortes C, Vapnik V. Support - Vector Networks. Machine Learning, 1995,20: 273 - 297

A SVD Method in Bilingual Information Filtering

Haiming Lu Jinhui Xu * Zengxiang Lu Yanda Li

Dept. of Automation *Dept. of Computer Tsinghua University Beijing 100084

Email: luhm @jerry. au. tsinghua. edu. cn

Abstract This paper introduces a SVD method in bilingual information filtering. It gives an uniform presentation to bilingual documents. Then any arithmetic used in monolingual information filtering can be easily used in bilingual information filtering. Using this method, we can compress the document vector and filter the noise. This method is used in personal information filtering on the Internet. We provide the WWW Bookmark Service. Through user's Bookmark, we can get user's preference and recommend interesting bilingual documents. According to user's feedback, we can improve the quality of information filtering.

Keywords Bilingual information filtering SVD Internet Bookmark Service