基于向量空间模型的文本过滤系统*

黄萱菁+, 夏迎炬, 吴立德

(复旦大学 计算机科学与工程系,上海 200433)

A Text Filtering System Based on Vector Space Model

HUANG Xuan-Jing+, XIA Ying-Ju, WU Li-De

(Department of Computer Science and Engineering, Fudan University, Shanghai 200433, China)

+ Corresponding author: Phn: 86-21-65642192, E-mail: xjhuang@fudan.edu.cn

http://www.fudan.edu.cn

Received 2001-09-14; Accepted 2002-04-10

Huang XJ, Xia YJ, Wu LD. A text filtering system based on vector space model. *Journal of Software*, 2003,14(3):435~442.

Abstract: Text filtering is the procedure of retrieving documents relevant to the requirements of specific users from a large-scale text data stream. First, the TREC (text retrieval conference) as well as its text filtering track are introduced, which is the most authoritative international evaluation conference on text retrieval, from the aspects of tasks, topics, corpus and evaluation metrics. Then a text filtering system based on vector space model is presented. This system is composed of two phases of training and adaptive filtering. During the training phase, feature selection and pseudo feedback are used to select the initial filtering profiles and thresholds. During the filtering phase, user feedback is utilized to modify the profiles and thresholds adaptively. This system took participate in the 9th Text Retrieval Conference in 2000, and ranked high among all the 15 systems from many countries. Good performance has been achieved, where the average precisions of adaptive and batch filtering are 26.5% and 31.7% respectively.

Key words: text retrieval; text filtering; text categorization; machine learning; vector space model

摘 要: 文本过滤是指从大量的文本数据流中寻找满足特定用户需求的文本的过程.首先从任务、测试主题、语料库和评测指标等方面介绍了文本检索领域最权威的国际评测会议——文本检索会议(TREC)及其中的文本过滤项目,然后详细地描述了基于向量空间模型的文本过滤系统.该系统由训练和自适应过滤两个阶段组成.在训练阶段,通过特征抽取和伪反馈建立初始的过滤模板,并设置初始阈值;在过滤阶段,则根据用户的反馈信息自适应地调整模板和阈值.该系统参加了2000年举行的第9次文本检索会议的评测,取得了很好的成绩,在来自多

* Supported by the National Natural Science Foundation of China under Grant Nos.69873011, 69935010, 60103014 (国家自然科学基金); the National High Technology Development 863 Program of China under Grant No.863-306-ZD02-02-4 (国家863高科技发展计划); the National High-Tech Research and Development Plan of China under Grant No. 2001AA114120 (国家高技术研究发展计划); the Science and Technology Promotion Foundation of Shanghai of China under Grant No.995115005 (上海市科学技术发展基金); the Science and Technology Foundation of Fudan University of China (复旦大学科学技术基金)

第一作者简介: 黄萱菁(1972－),女,浙江平阳人,博士,副教授,主要研究领域为大规模文本信息处理.

个国家的 15 个系统中名列前茅,其中自适应过滤和批过滤的平均准确率分别为 26.5%和 31.7%.

关键词: 文本检索;文本过滤;文本分类;机器学习;向量空间模型

中图法分类号: TP181 **文献标识码**: A

文本过滤是指从大量的文本数据流中寻找满足特定用户需求的文本的过程.预先给定一个用户需求和一个输入文本流,文本过滤系统必须首先根据用户需求建立一个初始的用户模板(称为 Profile),然后判断流中的每一个文本是否符合用户需求,并将符合用户需求的文本提交给用户,由用户对文本作是否符合其需求的评判,再根据评判结果自适应地修改用户模板,以更好地符合用户的需求.

文本过滤和文本检索有很大的相似之处,所不同的是文本检索有相对固定的文本库和千变万化的检索需求,而文本过滤则有着相对固定的用户需求和动态变化的文本流.可以说,文本过滤和文本检索是同一硬币的正反两面[1].

文本过滤对大规模文本信息处理具有很重要的意义,它可以应用在许多不同的领域,例如提供选择性信息服务的企事业单位、档案管理领域和终端用户等[2,3].虽然如此,但与文本检索相比,文本过滤开展得较晚,研究得也相对较少些.究其原因,是因为文本过滤要求有一个大规模的、真实而又权威的语料库,并且需要有完备、客观的人工评价结果,以进行反馈和自适应,同时可对不同的过滤方法进行比较.而由于缺乏必要的人力、物力,上述环境在实验室是非常难以模拟的.

1 TREC 及其文本过滤项目

1.1 文本检索会议

20 世纪 90 年代以来,情况有了彻底的改变,著名的文本检索会议[4](text retrieval conference,简称 TREC)以及主题检测和跟踪会议[5](topic detection and tracking,简称 TDT)都把文本过滤作为主要研究内容之一,这就在很大程度上促进了文本过滤的发展.下面将着重介绍文本检索会议及其在文本过滤方面所做的工作.

文本检索会议,是由美国国家标准技术局(National Institute of Standards and Technology,简称 NIST)和国防部高级研究计划局(Defense Advanced Research Projects Agency,简称 DARPA)组织召开的一年一度的国际会议,是文本检索领域最权威的国际会议之一,代表了当今世界文本检索领域的最高水平.

TREC 会议的宗旨主要有 3 条[4]:通过提供规范的大规模语料(GB 级)和对文本检索系统性能的客观、公正的评测,来促进技术的交流、发展和产业化;促进政府部门、学术界、工业界之间的交流和合作,加速技术的产业化;发展对文本检索系统的评测技术.

2000 年的 TREC-9(第 9 次文本检索会议)共有来自世界各地的 70 多个单位参加,包括许多著名的大学和公司,其中复旦大学和微软亚洲研究院是第一批来自中国大陆的参加单位.

每届文本检索会议都针对当前文本检索会议的最新热点,设置若干个评测任务.早期的任务是标准的文本检索,近年来,随着文本检索技术的不断发展和成熟,文本检索会议也逐渐把评测任务转移到更加新颖的研究方向上(称为项目).TREC-9 共设置了包括文本过滤在内的 7 个评测项目;TREC2001 设置了包括文本过滤在内的 6 个评测项目;TREC2002 则设置了包括文本过滤在内的 7 个评测项目.

1.2 文本过滤的任务定义

作为一个崭新的研究领域,文本过滤项目的任务定义开始时是逐渐演化的,难度越来越大,以更好地模拟真实环境.从 1997 年的 TREC-6 开始,文本过滤的主要任务逐渐固定下来.下面给出 TREC-9 至今文本过滤项目的任务定义[6]:给定一个主题描述(即用户需求),建立一个能从文本流中自动选择最相关文本的过滤模板(filtering profile).随着文本流的逐渐进入,过滤系统自动地接受或拒绝文本,并得到文本相关与否的反馈信息,根据反馈信息自适应地修正过滤模板.

文本过滤项目包含 3 个子任务.一个是被称为分流(routing)的子任务.其定义为:用户需求固定,提供对应于

该用户需求的训练文本集中的相关文本,从用户需求构造查询语句来查询测试文本集.另一个是批过滤(batch filtering):用户需求固定,提供对应于该用户需求的训练文本集中的相关文本,构造过滤系统,对测试文本集中的每一个文本作出接受或拒绝的决策.最重要的子任务是自适应过滤(adaptive filtering).它要求仅仅从主题描述出发,不提供或只提供很少的训练文本,逐一判断输入文本流中的文本是否相关.“接受”的文本,能得到用户的反馈信息,用以自适应地修正过滤模板.而被“拒绝”的文本是不提供反馈信息的.这是最接近真实环境,但也是最困难的子任务.

下文将以 TREC-9 为例,介绍文本过滤项目所用的语料库、用户需求(称为主题)、评价和实现方法.

1.3 语料库、主题和相关性评价

每次文本检索会议的过滤项目在语料库和用户需求上都略有不同.以前的几次文本过滤都是在新闻语料上进行的,而 TREC-9 则采用了医学文献语料库 OHSUMED,这是著名的 MEDLINE 医学文献库的一个子集,由 1988 年~1991 年的医学文摘组成,共含文本 348 566 篇,来自 270 种医学期刊,总容量为 400M 字节.其中 1987 年的文摘将作为训练语料,而 1988 年~1991 年的文摘将作为测试语料.

测试主题共有两类.主要的一类为“OHSU”类型,共有 63 个,由标题和描述两部分构成,例如:

〈title〉 60 year old menopausal woman without hormone replacement therapy

〈desc〉 Are there adverse effects on lipids when progesterone is given with estrogen replacement therapy

另外还有一类主题称为“MeSH”类型,由医学索引词汇及定义构成.对每个主题,TREC 都提供了全部文本的相关性评价,供训练和评测之用.相关文本在全部文本的比例中是非常稀少的,例如,“OHSU”主题平均 10 000 篇文本中只有 1.76 篇相关文本.由此带来的数据稀疏问题是文本过滤中的难点之一.

对于批过滤,允许系统使用训练集中全部文本的相关性评价.而对于自适应过滤,只允许使用少量的相关文本(2 篇~4 篇),其余文本都必须视为不知道是否相关.

1.4 评价方法

分流子任务根据平均非插值准确率(average uninterpolated precision)来评价,这是一个文本检索指标,具体可参见文献[6].批过滤和自适应过滤子任务则按如下的 Utility 和准确率来评价.

1.4.1 Utility

给定主题和文本,文本可能相关,也可能不相关;过滤系统可能检出该文本,也可能未检出.于是可以建立如下的四分表:

	相关	不相关
检出	R^+/A	N^+/B
未检出	R^-/C	N^-/D

检出相关文本和未检出不相关文本都是过滤正确的情况.而未检出相关文本意味着遗漏,检出不相关文本意味着错检.线性 Utility 函数对这 4 种情况赋以相应的权重:

$$\text{Utility} = A*R^+ + B*N^+ + C*R^- + D*N^- . \tag{1}$$

这里的 $R^+ / R^- / N^+ / N^-$ 指的是每个主题 4 种文本的数量.参数 $A/B/C/D$ 决定了每种情况的代价.显然,$A,D\geq 0,B,C\leq 0$.又由于检出相关文本是最重要的,所以 $A\geq D$.Utility 值越大,系统的过滤性能就越好.定义评价指标:

$$T9U = \begin{cases} 2*R^+ - N^+, \text{if } (2*R^+ - N^+) > \text{Min}U \\ \text{Min}U, \quad \text{otherwise} \end{cases} . \tag{2}$$

将全部主题的 Utility 数值进行平均,就得到全局的 Utility 数值.Min$U$ 是一个固定的下限,设为 4 年–800.定义这个下限是为了使某些过滤效果极差,从而 $T9U$ 很低的主题不会对全局性能造成太大的影响.

1.4.2 准确率

这个指标强调过滤的准确率,但要求检出的文本数不少于一个下限 Min$D$:

$$T9P = R^+ / \max(\text{Min}D, R^+ + N^+) . \tag{3}$$

这里,Min$D$ 为 4 年 50 篇,差不多相当于每月 1 篇.对每个主题的 $T9P$ 进行平均,就得到全局的平均准确率.

1.5 实现方法简介

从各个系统的实现方法上来看,大致有两类方法:一类是基于信息检索的方法,包括向量检索和概率检索;另一类是基于文本分类的方法,例如最近邻分类、神经网、Boosting Beyes 分类器、决策树、动态聚类和支撑向量机等[6].

2 文本过滤算法

2000 年,复旦大学作为第一批来自中国大陆的研究单位参加了 TREC-9,完成了文本过滤等 3 个项目,并取得了较好的成绩.下面将介绍我们提出的基于向量空间模型的自适应过滤机制和结果.由于对 TREC-9 而言,针对"OHSU"类主题,采用 $T9P$ 作为评价指标的自适应过滤子任务是最重要的,因此我们所提供的数据主要是在这种条件下产生的.

由于向量空间模型具有表示简洁和计算简便的特点,我们采用该模型来表示模板、主题和文本.关于向量空间模型的细节,可参看文献[7,8].

基于向量空间模型的文本过滤包括训练和过滤两个阶段.训练阶段的目的是根据给定的训练数据,生成初始的过滤模板,并决定初始的阈值.在自适应过滤阶段,对于文本流中的每篇文本,系统判断它是否和过滤模板相关,再根据用户的反馈信息,自动调整过滤模板和阈值,以获得最佳的过滤性能.

2.1 文本过滤的训练算法

2.1.1 训练算法的体系结构

图 1 说明了训练算法的体系结构.首先,我们将主题转变为向量形式,同时从正例文本和伪正例文本中抽取特征向量.而初始的模板则是主题向量、正例特征向量和伪正例特征向量的加权和.于是我们就可以计算初始向量和全部的训练样本之间的相似度,从而为每个主题选择最优的初始相似度阈值.

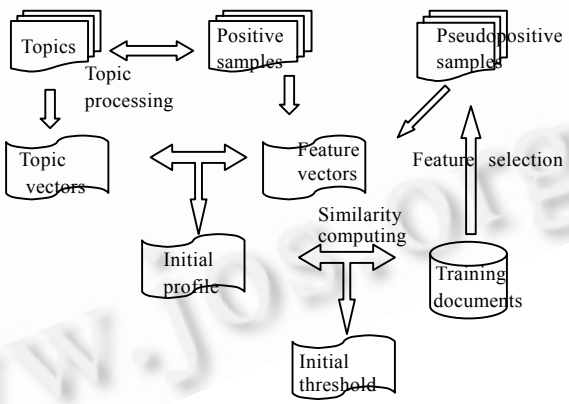

Fig.1 Architecture for the training algorithm in adaptive filtering

图 1 自适应过滤训练算法的体系结构

2.1.2 创建初始模板

初始模板向量是主题向量、由正例文本抽取的特征向量和由伪正例文本抽取的特征向量 3 个向量的加权和,权重分别为 α, β 和 γ,即

$$Pf_0(Q)=\alpha\cdot P_0(Q)+\beta\cdot P_1(Q)+\gamma\cdot P_2(Q), \tag{4}$$

其中 Q 表示主题,$Pf_0(Q)$是主题 Q 的初始模板向量,而 P_0,P_1 和 P_2 是它的 3 个分量.

对于主题向量 $P_0(Q)$,有 $P_0(Q)=(p_{01},p_{02},...,p_{0W})$.其中 W 表示词汇的总数,p_{0i} 是第 i 个词 w_i 的权重,采用 Smart 系统的 ltc 公式计算[9]:

$$p_{0i}=\begin{cases}\log(N/df(w_i)), & \text{if } w_i\in Q\\ 0, & \text{otherwise}\end{cases}. \tag{5}$$

这里,N 表示文本总数,而 $df(w_i)$表示词汇 w_i 的文本频数(即出现了 w_i 的文本的数量).若 w_i 未在主题 Q 中出现,则其权重为 0.

$P_1(Q)$是从正例文本中抽取的特征向量,$P_1(Q)=(p_{11},p_{12},...,p_{1W})$,$p_{1i}$ 是 w_i 的权重:

$$p_{1i}=\begin{cases}\log MI(w_i,rel(Q)), & \text{if } \log MI(w_i,rel(Q))\ge 3\\ 0, & \text{otherwise}\end{cases}. \tag{6}$$

之所以将许多词汇的权重赋为 0,是因为在大规模文本处理中,计算速度也是一个必须考虑的因素,若词汇的数量很多,在进行向量计算时就需要大量的时间.文本过滤的特征抽取,就是从全部可能的词汇(或称为特征)中抽取一个最优的特征子集,以降低向量空间维数,简化计算,防止过分拟合.而最优特征就是那些与相关文本集互信息量最大的词汇.词汇和相关文本集 $rel(Q)$之间的对数互信息量由下式计算:

$$\log MI(w_i,rel(Q))=\log\left(P(w_i\mid w_i\in rel(Q))\Big/P(w_i)\right). \tag{7}$$

$P_2(Q)$是从伪正例文本中抽取的特征向量.类似地,有 $P_2(Q)=(p_{21},p_{22},...,p_{2W})$,

$$p_{2i}=\begin{cases}\log MI(w_i,pseudo-rel(Q)), & \text{if } \log MI(w_i,pseudo-rel(Q))\ge 3\\ 0, & \text{otherwise}\end{cases}. \tag{8}$$

式中的$pseudo\text{-}rel(Q)$表示主题Q的伪正例文本,即与模板向量高度相似但又不是给定的正例文本的那些文本.这是由于对每个主题,我们只能得到少量的正例文本,因此我们需要加入伪反馈的功能,以便从训练文本中挖掘出更多的相关文本来补充正例文本.

2.1.3 设置初始阈值

相似度阈值一旦设立,那些与模板向量的相似度大于或等于阈值的文本就被认为是相关文本,而其他文本就被认为是不相关的.这样我们就可以计算在某个阈值水平下的性能评价指标(如 $T9U$ 或 $T9P$),选择能导致最佳性能的阈值作为初始阈值.若把 $T9P$ 看成是阈值 TH 的函数,对于初始阈值 $TH(0)$,则有

$$TH(0)=\underset{TH}{\mathrm{ARGMAX}}(T9P(TH)). \tag{9}$$

而模板向量和训练文本之间的相似度采用余弦公式[7]获得:

$$Sim(d,pf)=\sum_k d_k * pf_k\Big/\sqrt{(\sum_k d_k^2)(\sum_k pf_k^2)}. \tag{10}$$

这里的 pf 表示模板向量,d 表示文本.d_k 是 d 中第 k 个词的权重,$d_k=1+\log tf_k$,而 tf_k 是 d 中第 k 个词的频率.

2.2 文本过滤的自适应算法

2.2.1 自适应算法的体系结构

当初始的模板向量建立,并且阈值也已设置好之后,文本过滤的过程就是自适应地修改模板向量和阈值,使得过滤系统的性能不断提高的过程.图 2 说明了自适应算法的体系结构.

对文本流中的每篇文本,我们都可计算它和某个主题的模板向量的相似度.若相似度超过阈值,就被认为是相关文本.然后由用户判断这篇文本是否真正与主题相关.根据不同的结果再相应地修改模板向量或调整阈值.

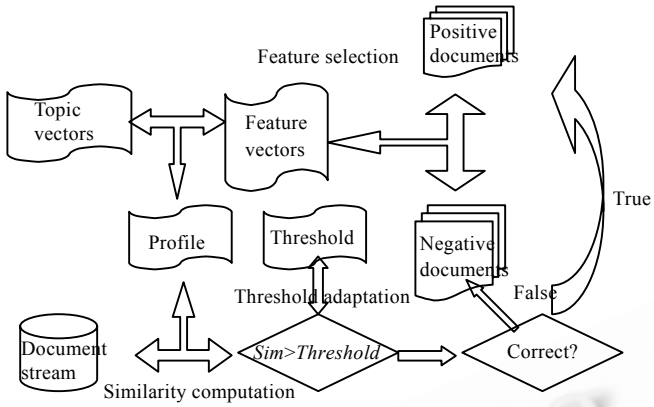

Fig.2 Architecture for the adaptation in adaptive filtering

图 2 自适应算法的体系结构

2.2.2 阈值的调整

由于真实文本流中相关文本的比例是很低的,因此我们并不是每次检出文章后就进行自适应,而是在经过一个固定长度的时段后才调整阈值.提高阈值的目的是检出较少的文本,从而提高准确率;而降低阈值的目的是检出较多的文本,从而提高召回率.用 M_0 表示期望在一个时段内检出的文本数量,用 T 表示第 T 个时段,对每个主题,我们定义:

$cor(T)$:在 T 时段正确检出的文本数　　$rtv(T)$:在 T 时段检出的文本数

$COR(T)$:到 T 时段为止正确检出的文本数　　$RTV(T)$:到 T 时段为止检出的文本数

$M(T)$:到 T 时段为止必须检出的文本数　　$TH(T)$:T 时段的相似度阈值

显然有 $$COR(T)=\sum_{t=1}^{T} cor(t)\ ,\ RTV(T)=\sum_{t=1}^{T} rtv(t)\ ,\ M(T)=M_0 * T\ .$$

我们提出了如下的阈值调整算法:

(1) 若 $cor(T)<rtv(T)*20\%$,且 $rtv(T)>\max(M_0,4)$,则 $TH(T+1)=TH(T)*1.2$.即如果准确率过低,而检出的文本又不太少,则迅速提高阈值.

(2) 若 $rtv(T)>M_0$ 且 $RTV(T)>M(T)$,则 $TH(T+1)=TH(T)*1.1$.即如果检出的文本多于必需的,则提高阈值.

(3) 若 $rtv(T)<M_0$ 且 $RTV(T)<M(T)$,则 $TH(T+1)=TH(T)*0.9$,即如果检出的文本少于必需的,则降低阈值.

2.2.3 模板的修改

一旦检出的文本被用户判断为相关文本,我们就将它加入到正例文本集合中,否则就加入到反例文本集合中.在调整模板向量时,我们从正例文本和反例文本中抽取出特征向量.于是新的模板向量就是主题向量、由正例文本抽取的特征向量和由反例文本抽取的特征向量 3 个向量的加权和,权重分别为 α', β' 和 γ',即

$$Pf'(Q)=\alpha'\cdot P_0(Q)+\beta'\cdot P_1(Q)+\gamma'\cdot P_3(Q). \tag{11}$$

式中的 $P_3(Q)$是由反例文本抽取的特征向量, $P_3(Q)=(p_{31}, p_{32},..., p_{3W})$,

$$p_{3i}=\begin{cases}\log MI(w_i,irrel(Q)), & \text{if } \log MI(w_i,irrel(Q))\geq 3\\ 0, & \text{otherwise}\end{cases}. \tag{12}$$

式中的 $irrel(Q)$表示主题 Q 的反例文本.

2.3 结果及分析

文本过滤项目的整体评测结果见表 1[6].

Table 1 Best text filtering systems

表 1 最好的一批文本过滤系统

Routing P@50		Batch filtering *T9P*		Adaptive filtering *T9P*	
ICDC	37.0	Fudan	31.7	Microsoft	29.4
Microsoft	33.6	Microsoft	30.5	CMU-LT1	27.9
Nijmegen	28.2	CMU-Y	26.1	Fudan	26.5
OHSU topics Filtering runs optimized for *T9P* Best runs from best 3 groups					

为了进一步说明自适应过滤的性能,我们还给出了 *T9P*,Micro*F* 和 Macro*F* 随时间变化的曲线(Micro*F* 和 Macro*F* 分别表示微平均和宏平均的 *F* 值,它们是准确率和召回率的一种平均,其具体定义可参见文献[10]),如图 3 所示.

图中的横轴表示时间,纵轴表示 3 种评价指标.自适应过滤开始于 1988 年,结束于 1991 年.从图中我们可以看出,开始时过滤系统的性能还不是很好,几个评价指标都处在一个较低的水平,随着时间的推移,各种评价指标一致地上升.这说明系统确实具有较好的自适应能力,能有效地从用户的相关性评价获取信息,提高系统的性能.

图 4 同样也证实了这一点.图中的横轴是 63 个主题,按批过滤的准确率 *T9P* 从大到小的顺序排列.纵轴则给出了每个主题自适应过滤和批过滤的 *T9P* 数值.批过滤每个主题平均提供了 10 篇相关文本,而自适应过滤只提供了 2 篇.此外,我们还给出了在每个主题只提供 2 篇相关文本,且不进行自适应的情况下的 *T9P* 数值.

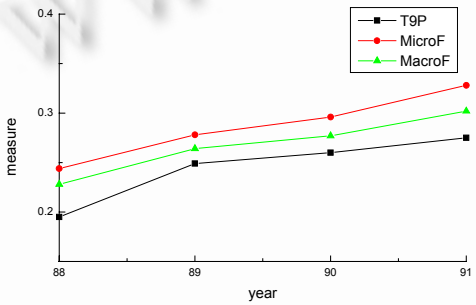

Fig.3 Adaptive filtering by year

图 3 过滤性能和时间的关系

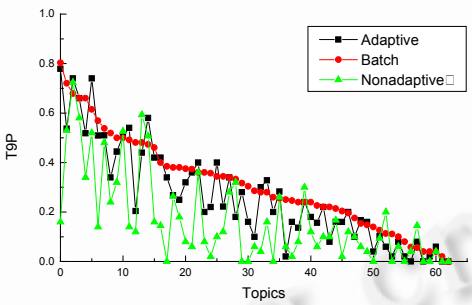

Fig.4 Adaptive filtering vs. batch filtering

图 4 自适应过滤和批过滤的性能比较

从图中可以发现,与批过滤相比,自适应过滤的性能下降得并不是很大,两条曲线非常接近.事实上,两者的平均数值分别是 31.7%和 26.5%,下降幅度仅为 16.4%.相比之下,在不进行自适应的情况下,大多数主题的 *T9P* 均有很大幅度的下降;且平均 *T9P* 仅为 17.5%,下降了 45.8%.这就充分说明了自适应的作用.

3 结 论

本文介绍了 2000 年召开的第 9 次文本检索会议的文本过滤项目,并详细地介绍了我们提出的基于向量空间模型的文本过滤系统.该系统包括训练和过滤两个阶段,在来自多个国家的 15 个研究小组中取得了自适应过滤的第 3 名和批过滤的第 1 名.

本文所介绍的文本过滤系统就其本质而言,采用的仍然是浅层的统计方法.今后我们将尝试较深层次的处理,例如语义分析等.未来的工作是在大规模的中文语料库上实现上述的文本过滤系统.

总之,经过多年的实践,TREC 已经建立了在文本检索会议的国际权威地位,吸引了世界各地越来越多的高水平的研究机构,也发展了一套较为成熟的评测方法.目前,国内对 TREC 和文本过滤感兴趣的研究单位也越来越多,相信通过大家的努力,国内的文本过滤水平一定会更上一层楼.

References:

[1] Belkin N, Croft WB. Information filtering and information retrieval, two sides of the same coin. Communications of the ACM, 1992,33(12):29~38.

[2] Daniel EO. The Internet, Intranet, and the AI renaissance. Computer, 1997,30(1):71~78.

[3] David DL. The TREC-4 filtering track. In: Harman DK, ed. Proceeding of the 4th Text Retrieval Conference (TREC-4). Gaithersburg: NIST Special Publication, 1995. 165~180.

[4] Voorhees EM, Harman DK. Overview of the 9th text retrieval conference (TREC-9). In: Voorhees EM, Harman DK, eds. Proceedings of the 9th Text Retrieval Conference (TREC-9). Gaithersburg: NIST Special Publication, 2000. 1~14.

[5] Charles LW. Topic detection & tracking (TDT) overview & perspective. In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop. 1998. http://www.nist.gov/speech/publications/darpa98/index.htm, Lansdowne.

[6] Robertson S, Hull DA. The TREC-9 filtering track final report. In: Voorhees EM, Harman DK, eds. Proceedings of the 9th Text Retrieval Conference (TREC-9). Gaithersburg: NIST Special Publication, 2001. 25~40.

[7] Salton G. Developments in automatic text retrieval. Science, 1991,253(5023):974~979.

[8] Wu LD, Huang XJ. Large-Scale Chinese Text Processing. Shanghai: Fudan University Press, 1997. 102~118 (in Chinese).

[9] Buckley C, Salton G, Allan J. Automatic retrieval with locality information using SMART. In: Harman DK, ed. Proceedings of the 1st Text REtrieval Conference (TREC-1). Gaithersburg: NIST Special Publication, 1992. 59~72.

[10] Huang XJ, Wu LD, Ishizaki Hiroyuki, Xu GW. Language independent text categorization. Journal of Chinese Information Processing, 2000,14(6):1~7 (in Chinese with English Abstract).

附中文参考文献:

[8] 吴立德,黄萱菁.大规模中文文本处理.上海:复旦大学出版社,1997.102~118.

[10] 黄萱菁,吴立德,石崎洋之,徐国伟.独立于语种的文本分类方法.中文信息学报,2000,14(6):1~7.