从文档集推导 html 标签影响因子的算法*)

邓剑勋　邢永康
(重庆大学计算机学院　重庆 400044)

摘　要　在 Web 文档中,同一个关键词处在不同 html 标签中,其对中心思想影响程度各不相同。选择合适的标签影响因子,对于构建文档的数学模型至关重要。本文在总结前人研究基础之上,提出了一种新的推导算法,该算法提出了 ttf(标引词标签频率)和 itf(逆标签频率)等定义,构造出行序为标签、列序为关键词的文档矩阵。从中抽取每个文档的某一特定行向量组构成一个新的向量集合,根据这个新集合中各个向量到质心的平均距离,就能得出该特定行向量集合所代表的标签的影响因子(针对训练文档集合)。如果训练文档集合的容量放大到足够,就可以近似认为这个影响因子具有一般意义。通过试验验证,推导出的影响因子作用于新的文档集合的时候,在一定程度上改善了检索的性能。

关键词　ttf,itf,规范化因子,质心,平均距离,标签影响因子向量

Arithmetic of Deriving Html Tags Influence Factor from Docment Collections

DENG Jian-Xun　XING Yong-Kang
(Computer Science Department of Chongqing University, Chongqing 400044)

Abstract　In html documents , one kind of keyword may have different influence factor to main idea, because it lay in different html tag . So it's important to choose a suitable common influence factor in setting up a math model of html document . This paper, based on the recently research, brings out a new deriving arithmetic. The arithmetic , ground on some new concepts, such as ttf(Term Frequency in Tag),itf(Inverse Tag Frequency), transform one document to a matrix which row represent html tags and column represent keywords . Use certain row (as a row vector) in every document to form a new vector list , then calculate the average distance between every vector to Centroid in the list . By the averay distance we can get the tag's influence factor(to the documents aggregate we used) . If the documents aggregate is big enough , then the influence factor we get is approximately be regard as the common influence factor. Apply the result in the new documents aggregate , we find searching is effective than before.

Keywords　ttf,itf,Standardization factor,Centroid,Average distance,Tags influence factor

1　引言

html 文档不同于普通文本文档,其中使用了大量的标签,在基于 VSM(向量空间模型)[1]的聚类和分类算法中,忽略这些标签对结果的影响,将会影响到检索算法的评价标准。文[8]中把标签分成了 4 类,按类别由领域专家赋值,而文[9]一文中只把标签分成了三类。这些方法主观性太强。本文认为特定文档本身已包含了很多区分标签影响度的信息,且不同文档中同种标签的影响度也存在差异,基于此提出了一种新的算法,该算法能从已知文档集合中抽取出具有一定代表性的标签影响因子向量,如果集合足够大,得到的影响因子向量应具有一般性,可以推广使用。本文使用基于 VSM[3]的检索算法,对该算法进行验证得出结论:本文提出的算法求出的标签影响因子,应用到实际中后,在一定程度上改善了 VSM 检索算法的性能。

2　标签影响因子算法

标签影响因子算法,本质就是从大量已知文档集合中提取出具有一般性推广价值的标签影响因子。其主要思想就是:

任意一个 Web 文档,html 标签影响因子的信息应该包含在这个文档所承载的信息中。这也就是说,单个文档本身包含了相当的标签影响度信息。

同样一个 html 标签,在不同的 Web 文档中,其相对主题的重要程度实际上是有所差别的。对大容量文档集合进行综合分析,能得出具有一般性的标签影响因子度量(在集合足够大的情况下)。

鉴于此,我们给出了一种方法,该方法能针对某一文档集合得出一个影响因子序列(序列中为各个 html 标签对表达文档主题的影响因子,结果针对该集合)。那么,如果给出一定容量的训练文档集合,把每个文档(doc_k)表示成行序为 html 标签序列($Tag=\{tag_1,tag_2,\cdots,tag_n\}$),列序列为分词字典中标引词序列($Dic=\{dic_1,dic_2,\cdots,dic_s\}$)的一个矩阵,

$$doc_k=\begin{array}{c} \cdots\cdots \quad dic_1 \quad dic_2 \quad \cdots \quad dic_s \\ \begin{matrix} tag_1 \\ tag_2 \\ \cdots \\ tag_n \end{matrix} \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1s} \\ p_{21} & p_{22} & \cdots & p_{2s} \\ \cdots & \cdots & \cdots & \cdots \\ p_{n1} & p_{n2} & \cdots & p_{ns} \end{pmatrix} \end{array}$$

*)基金项目:本研究得到国家自然科学基金青年基金资助(编号:60403009)。**邓剑勋**　硕士研究生,助理工程师,主要研究方向为人工智能、信息检索;**邢永康**　博士,副教授,主要研究方向为人工智能、信息检索和数据挖掘以及 Bayes 网。

其中 doc_k 中权重 p_{ij} 计算方法如下：

$$p_{ij}=ttf\times itf/\sqrt{w_1^2+w_2^2+\cdots+w_n^2} \tag{1}$$

下面简单介绍一下公式(1)中的几个定义。

(1)标引词标签频率 ttf。ttf=该标签内该标引词出现次数/该标签内中的标引词个数 (2)

(2)逆标签频率 itf。标签频率度量的是文档中包含了该项标引词的标签个数，用 tf 表示。有如下换算公式：

$$itf=1/tf \tag{3}$$

(3)规范化因子。不同标签内字符串长度不同，因此对项频有很大影响。为了抵消这种由篇幅带来的影响，需要对项影响因子进行规范化处理。我们可以采用余弦规范[6]来处理影响因子。

$$\sqrt{w_1^2+w_2^2+\cdots+w_n^2} \tag{4}$$

其中 w_i 是第 i 个项在字符串中的 $ttf\times itf$ 值。接下来我们的工作就是获得标签影响因子向量：

每一个文档矩阵中的一个行向量，都体现了对应标签在该文档中的特征程度。由于在不同的文档中，同一个标签的专指程度可能会根据语境有所不同，因此，如果对大量的训练文档集合进行处理，把每一个 doc_k 中 tag_i 对应的这一行作为一个行向量 $row_k[k$ 是文档号$]$，把这些行向量归并为一个集合，求出该集合的质心，然后用各个 tag_i 行向量到质心的平均距离，作为标签影响因子的一个近似拟合。那么，当训练文档集合的容量足够大的时候，我们可以近似地认为，这个值就是具有一般性的标签影响因子。根据这个理论，我们有下面的算法。

算法 获得标签影响因子向量 $TagPowArr$

输入：文档矩阵集合 $Doc=\{doc_1,doc_2,\cdots,doc_k\}$，一个标引词字典 $Dic=\{dic_1,dic_2,\cdots,dic_s\}$ 以及一个标签集合 $Tag=\{tag_1,tag_2,\cdots,tag_n\}$

输出：一个标引词影响因子向量 $TagPowArr$

步骤：1)遍历每个文档矩阵的第 $i[i=1\ to\ n]$ 个行向量，构成一个行向量集合 Arr_i，求出行向量集合的质心[6]

Arr_i 内的向量可以表示为：

$$A_m=(p_{i1[m]},p_{i2[m]},\cdots,p_{is[m]}) \tag{5}$$

其中 m 为 1 到 k 之间的一个数，该向量表征的含义是序号为 i 的标签在 m 文档中的重要度量。故质心是：

$$C_i=(\frac{1}{k}\sum_{m=1}^{k}p_{i1[m]},\frac{1}{k}\sum_{m=1}^{k}p_{i2[m]},\cdots,\frac{1}{k}\sum_{m=1}^{k}p_{is[m]}) \tag{6}$$

也可以表示为：

$C_i=(c_1,c_2,\cdots,c_s)$

2)求出行向量集合 Arr_i 中每个行向量到质心 C_i 的距离

按照如下计算式计算向量到质心的距离：

向量依然使用 $A_m=(p_{i1[m]},p_{i2[m]},p_{is[m]})$ 表示方法，含义同上。

那么从 m 文档中取出的向量和 Arr_i 质心的距离是

$$d_m=[\sum_{f=1}^{s}(p_{if[m]}-c_f)^2]^{1/2} \tag{7}$$

求解过程中我们使用的是欧氏距离[6]。

3)求出距离的平均值

$$\overline{d}=(\sum_{m=1}^{k}d_m)/k \tag{8}$$

作为 $TagPowArr$ 的第 i 个分量值。

4)不断重复 1、2、3 步(注意 i 值从 1 递增到 n)，可以为 $TagPowArr$ 中的每一个元素赋值。

5)输出标签影响因子向量组 $TagPowArr$。

3 实验与结论

在基于 VSM 的向量空间分类中，文档往往通过 tf/idf 方式[3]表示成向量。对于一个文档集合，我们通常表示成如下形式：

$$\begin{array}{c} \cdots\cdots \quad key_1 \quad key_2 \quad \cdots \quad key_m \\ \left.\begin{array}{l} doc_1 \\ doc_2 \\ \cdots \\ doc_n \end{array}\right\{ \begin{array}{cccc} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ p_{n1} & p_{n2} & \cdots & p_{nm} \end{array}\right\} \end{array} \tag{9}$$

其中矩阵的分量一般采用词频和逆文档频率乘积的方式来表示：

$$p_{ij}=tf_{ij}*idf_{ij} \tag{10}$$

传统的 VSM 通常采用如下计算表达式来求解词频：

$$tf_{ij}=\frac{\text{文档 } i \text{ 中关键词 } j \text{ 出现的次数}}{\text{文档 } i \text{ 中关键词的个数}} \tag{11}$$

采用我们前面得出的 $TagPowArr$ 向量，我们对上面这个公式进行修正。

$$tf_{ij}=\{\sum_{m=1}^{s}\frac{\text{文档 } i \text{ 中标签 } m \text{ 内关键词 } j \text{ 出现的次数}}{\text{文档 } i \text{ 中标签 } m \text{ 内关键词的个数}}*TagPowArr[m]\}/s \tag{12}$$

这样我们就得到一个改进的文档表示方法。把这个改进的表示方法应用到 VSM 模型中，可以验证修正前后 VSM 模型的检索指标，从而衡量我们算法的有效性。

我们选择了 350 篇关于计算机方面的 HTML 文档，并将其分为 5 类：网络通信、数据库、程序设计、人工智能和项目管理，每个类别 70 篇文档。将 350 篇 HTML 文档的数据集合分为训练集(250 篇) 和测试集(100 篇)，使用基于 VSM 的检索算法通过不断的学习和测试，就可以得到一组结果。

使用查准率(precision)[7]和查全率(recall)[7]来评价检索的质量。表 1 表示了修正前后对检索质量的影响。

表 1

	修正前		修正后	
	precision	recall	precision	recall
网络通信	62.15%	66.28%	63.17%	68.06%
数据库	51.28%	57.14%	61.25%	62.26%
程序设计	52.43%	59.48%	50.60%	60.27%
人工智能	63.23%	66.27%	67.19%	66.16%
项目管理	58.87%	59.29%	65.26%	63.31%

结论 Internet 上的资源大多以 HTML 格式的文件存在，如何准确和高效地获取满足用户需求的信息成为信息检索领域一个重要的研究内容。结合 HTML 文件的结构特征，基于标签影响因子的文档表示方法，在基于 VSM[2]的检索算法上应用表明，该标签影响因子算法在实际的文本检索中取得良好的效果。但是该算法还有一定的局限性，因为训练文档数量不可能无限增加，所以其代表整体标签影响因子的可信度还不是 1，而是介于 0 和 1 之间的一个小数。在今后的研究中，力争引入标签影响因子代表性可信度的概念，使其更适于实际情况。同时，我们求出来的标签影响因子，在修正 VSM 模型的时候，目前采用的是加权求平均的方式，如公式(12)，也可以采用最大值法，如：

$$tf_{ij}=\max_{m=1}^{s}(\frac{\text{文档 } i \text{ 中标签 } m \text{ 内关键词 } j \text{ 出现的次数}}{\text{文档 } i \text{ 中标签 } m \text{ 内关键词的个数}}*TagPowArr[m])$$

基于粗糙集的个性化 Web 搜索系统

时 雷[1] 席 磊[1] 段其国[2]

(河南农业大学信息与管理科学学院 郑州 450002)[1] (同济大学计算机科学与工程系 上海 201804)[2]

摘 要 本文提出了一种基于粗糙集理论的个性化 Web 搜索系统。用户偏好文件中对关键字进行分组以表示用户兴趣类别。利用粗糙集理论处理自然语言的内在含糊性,根据用户偏好文件对查询条件进行扩展。搜索组件使用扩展后的查询条件搜索相关信息。为了进一步排除不相关信息,排序组件计算查询条件和搜索结果之间的相似程度,根据计算值对搜索结果进行排序。与传统搜索引擎进行了比较,实验结果表明,该系统有效地提高了搜索结果的精度,满足了用户的个性化需求。

关键词 Web 检索,粗糙集,个性化

Personalized Web Search System Based on Rough Sets

SHI Lei[1] XI Lei[1] DUAN Qi-Guo[2]

(College of Information and Management Science, Henan Agricultural University, Zhengzhou 450002)[1]

(Department of Computer Science and Technology, Tongji University, Shanghai 201804)[2]

Abstract In this paper, a novel rough set based approach is proposed to create a personalized Web search system. Firstly, user profiles which consist of categories of user's interests by grouping related keywords are designed. Rough set theory is used to deal with inherent ambiguities of natural language and refine query according to user profiles. Then refined query is submitted to search component. To further filter out irrelevant documents for the user, retrieved results are re-ranked according to rough similarity measures between refined query and documents by ranking component. Experiments compared with traditional search engine are presented and experimental results indicate the precision of Web retrieval is greatly improved and system are suitable for individual usage.

Keywords Web retrieval, Rough sets, Personalization

1 引言

近年来随着网络和信息技术的发展,Web 上的信息量迅速增加。如何快速、准确地从浩瀚的信息资源中寻找到所需信息已经成为信息检索、Web 挖掘等领域的重要研究内容。Web 搜索引擎在一定程度上解决了这一问题,但传统的搜索引擎没有考虑用户的兴趣爱好。用户在使用搜索引擎时,常常是搜索到的网页成千上万,而真正需要的信息却又寥寥无几。在这种背景下,个性化的 Web 信息检索技术引起了越来越多的研究者的关注。个性化的 Web 检索有效地利用了用户的兴趣和偏好,可以帮助用户更加精确地检索信息。

随着个性化的 Web 检索研究的开展,研究成果和应用系统纷纷出现,如 Syskill&Webertt、WebWatcher 等[1~4]。这些系统大多采用了传统的信息检索模型,即布尔模型、向量模型和概率模型。然而,自然语言文档在本质上是基于上下文相关的,因此严格的二值逻辑无法处理自然语言中的细微差别。粗糙集理论是处理不精确、不确定和模糊信息的有力工具,它提供了一种近似地表示概念的方法。因为孤立地考虑文档中的词是无法对文档进行准确分类的,所以可以利用粗糙集对文档空间和查询项空间进行扩展,使其包括概念上相关的其它词。

本文提出了一种基于粗糙集的个性化 Web 检索系统。该系统利用粗糙集理论处理自然语言的内在含糊性,根据用户偏好对查询条件进行扩展。搜索组件使用扩展后的查询条件搜索相关信息。为了进一步排除不相关信息,排序组件计算查询条件和搜索结果之间的相似程度,根据计算值对搜索结果进行排序。

2 粗糙集

粗糙集理论是一种处理不完全或者不精确信息的工具[5,6],它已经在数据挖掘、机器学习和模式识别等领域中得到了广泛的应用。

时 雷 硕士,助教,研究方向:模式识别,数据挖掘;席 磊 硕士,讲师,研究方向:模式识别;段其国 博士生,研究方向:数据挖掘。

但是这几种修正方法都不一定是最科学的,如何根据具体的应用环境,让修正方式更加合理,是我们下一步的一个努力方向;进一步地,如何把修正矩阵从 VSM 推广到概率模型[5]和信度网[4]中,是我们需要进一步研究的重点。

参 考 文 献

1 Turney P D. Similarity of semantic relations. Computational Linguistics, 2006, 32 (3): 379~416

2 Zhu Y W, Hu Y M. Enhancing search performance on gnutella-like P2P systems. IEEE Transactions on Parallel and Distributed Systems, 2006, 17 (12): 1482~1495

3 Wang H M, Rajman M, Guo Y, et al. New PR-combining TFIDF with Pagerank. Artificial Neural Networks-icann, PT2 Lecture Notes in Comuter Science, 2006, 4132: 932~942

4 沈一栋,邢永康. 一种新的知识表达模型——信度网[J]. 计算机科学, 2000,9(27):40~43

5 邢永康,马少平. 信息检索的概率模型[J]. 计算机科学,2003,30(8):13~17

6 孙建军,成颖,等. 信息检索技术. 科学出版社, 204. 56~58

7 许忠锡. 查全率和查准率关系辨析. 上海高校图书情报研究, 2004,4:21~23

8 宋斌,方小璐. 基于网页特征的 TFIDF 改进算法. 微机算机应用,2002, 23(1):18~20

9 许建潮,胡明. 中文 Web 文本的特征获取与分类. 计算机工程, 2005,31(8):24~25,39