文章编号: 1673- 0062(2010) 01- 0039- 04

基于查询与内容的文档表示模型

阳小华, 周 座

(南华大学 计算机科学与技术学院, 湖南 衡阳 421001)

摘 要: 在信息检索中, 文档表示模型的优劣是影响检索性能的重要因素之一. 根据全信息理论, 认识论信息是语法信息、语义信息和语用信息的三位一体. 当前主流的文档表示模型主要利用语法和语义信息, 造成语用信息的缺失, 成为改善检索性能的瓶颈. 该文提出了一种整合用户查询行为与文档内容的文档表示模型, 将用户隐性反馈的语用信息和文档自身的语义、语法信息相结合, 动态调整索引库关键词权重, 从而提高信息检索的查全率和查准率.

关键词: 文档表示模型; 用户查询日志; 隐性反馈
中图分类号: TP391 文献标识码: A

Document Representation Model Based on Query and Content

YANG Xiao-hua, ZHOU Zuo

(School of Computer and Technology, University of South China, Hengyang, Hunan 421001, China)

Abstract: In information retrieval, the quality of a document representation model is one of the important factors which affect retrieval performance. According to the comprehensive information theory, epistemology information is the trinity of syntactic information, semantic information and pragmatic information. The mainstream of document representation models at present primarily utilize syntactic and semantic information while are devoid of pragmatic information, which is the bottle- neck of retrieval performance improving. In this thesis, we present a document representation model based on users' query behavior and documents' content, in which the pragmatic information from users' implicit feedback and the semantic and syntactic information from documents is integrated to dynamically regulate the key- weight of index database. This model can consequently improve recall and precision rate in information retrieval.

Key words: document representation model; user query log; implicit feedback

收稿日期: 2010- 01- 25
基金项目: 湖南省科技厅基金资助项目 (2006GK3086)
作者简介: 阳小华 (1963-), 男, 湖南衡阳人, 南华大学计算机科学与技术学院教授, 博士生导师. 主要研究方向: 智能信息系统与知识管理、软件工程.

0 引言

信息检索是一门研究信息的获取、表示、存储、组织和访问的学科, 它是一个从文档集合(Collection)中返回满足用户需求的相关信息的过程. 文档表示模型的优劣直接影响到信息检索的性能, 对于改善检索的查全率和查准率至关重要. 人们提出了布尔模型、向量空间模型、概率检索模型、语言模型等多种文档表示模型, 它们基本上都是基于原始文档的内容表示文档的主题信息, 这类模型的共同点是仅利用语法信息或语义信息来表示文档, 造成了语用信息的缺失. 根据钟义信教授的全信息理论[1], 认识论信息是语法信息、语义信息和语用信息的三位一体, 即信息的外在形式、内在含义和效用价值组成了一个完整的系统, 全面地理解信息应当利用全信息理论. 因此, 文档表示模型对语用信息的忽略和缺失势必在一定程度上影响文档表示的信度和效度, 成为改善检索性能的瓶颈. 充分地挖掘语用信息, 将其引入到文档表示模型中来, 必将有效地提高信息检索的性能. 人们在这方面开展了一系列工作, 如文献[2-5]所阐述的文档表示方法. 在文献[2]中, B. Poblete等提出了一种利用隐性用户反馈信息和查询集来组织Web文档的方法, 通过引入用户反馈将语用信息引入到文档表示中, 从而改善了检索质量. 本文在文献[2]的基础上, 尝试将用户查询集所含的语用信息与文档作者的语法与语义相结合, 提出了一种新的文档表示模型, 为进一步提高检索质量奠定基础.

1 基于用户反馈与查询集的文档表示模型

文献[2]提出了一种从搜索引擎的用户查询日志中挖掘出查询词或共现查询词并将其作为文档特征词构建文档表示模型的方法, 特征词的权重为引发Web文档被用户点击的查询词出现的频率. 其基本思路是: 在Web信息检索中, 当大量用户用某些相同或相似的查询词进行检索, 并且依据摘要信息对检索结果中的部分文档发生点击行为时, 则可以断定用户认为该文档与查询词具有相关性, 换言之, 该查询词较好地描述了该文档的主题信息. 引发某一文档被点击的查询词出现的频率越高, 则该查询词与文档内容的关联性越强, 应当赋予更大的权重. 为了界定每个查询词的区分度, 引入逆查询串频率, 在用户查询日志中抽取出所有的查询串(或查询语句), 某一查询词在越多的查询串中出现, 则该查询词对查询串的区分度越小, 也就对文档的区分度越小.

该模型中文档的表示基于传统的向量空间模型方法, 所不同的是, 该文档表示模型中的特征词与文档内容无关, 仅与引发该文档被点击的用户查询词相关. 实现步骤是: 首先, 从长期积累的用户查询日志中挖掘出所有的查询词(经过消除停用词等预处理), 统计词频及词与文档的关联程度(点击频率), 然后利用带权查询词作为特征词将文档表示为特征向量, 构建基于用户反馈与查询集的新向量空间模型. 该模型对于文档内容来说是无知的(content-ignorant), 它认为从查询中挖掘到的高频词集比从文档全文中抽取的词集更能反映用户的意图, 更能准确地从用户的角度理解文档的主题信息. 图1展示了用查询集表示文档的一个简单的例子, 给出了5组查询及其所含的查询词、用户通过查询串检索并点击各个文档的次数. 通过对这些信息进行处理来构造文档表示形式.

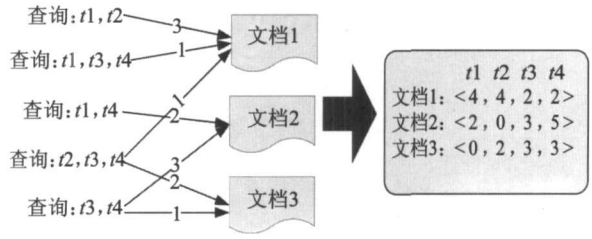

图1 用查询词表示文档的一个例子, 未经过标准化处理

Fig. 1 Example of the query document representation, without normalization

该模型可形式化定义为[2]:

$\{d_1, d_2, \cdots, d_n\}$表示一个文档集, V表示从访问日志L的所有查询串中抽取出的查询词表, $t_1, t_2, \cdots, t_m$表示查询词表V中的词语序列, $Q(d_i)$表示L中出现的各个查询串, 且由此查询串至少引发用户点击文档d_i一次, $Q(d_i)$中t_j的词频为包含t_j的查询串访问d_i的总次数. d_i的向量表示定义为:

$$\vec{d_i} = < C_{i1}, C_{i2}, \cdots, C_{im} >$$

$$C_{ij} = tf * idf(t_j, Q(d_i)) \quad (1)$$

$tf * idf(t_j, Q(d_i))$是分配给查询串$Q(d_i)$中的查询词t_j的$tf * idf$权重. 该模型的主要缺陷是将各个查询词看成是独立的, 查询词之间互不相

关, 而事实并非如此, 因为共现查询词的重复出现必然表示其词义在概念上相互关联. 文献[2]提出的改进模型的构造方法是: 先从用户查询日志中抽取出每个被点击文档对应的共现查询词, 然后为文档表示模型创建文档的特征向量. 特征空间的每个维度用从查询日志中挖掘出的一对共现词表示, 每对共现词是一个不能分割的单元, 被点击文档的每个特征维度的权重用其所对应的共现查询词在所有查询中出现的频率来表示.

文献[2]提出的两个文档表示模型将用户隐性反馈信息应用到信息检索的文档表示中来, 将语用信息引入文档表示模型, 并在一定程度上减少了文档的特征空间的维度, 但仍然存在一定的局限性. 首先, 该模型仅考虑与已有用户查询相关联的文档, 而添加到服务器端的新文档尚没有查询与之关联则不能被表示和检索. 该模型的实现必须建立在用户查询长期积累的基础之上, 而且基本上要求大量用户具有共同的兴趣.

更为重要的是, 该模型仅仅使用查询词和部分用户反馈信息来表示文档的主题信息, 完全摒弃了文档本身的内容所携带的信息. 即单纯从用户的角度来理解文档的主题信息, 彻底抛弃了文档创建者或系统的观点, 这显然是不合理的. 因为从理论上来看, 文档本身携带的信息往往更真实、全面, 而且从文档创建者的角度理解的文档主题信息也同样非常重要, 并非没有合理性, 但是该模型却彻底丢弃了这部分信息, 这势必在一定程度上影响检索性能, 限制检索的查全率和查准率. 另外, 从收集的数据来看, 查询样本空间的维度要小于文档空间的维度, 如果仅限于利用查询空间的信息来构建文档表示模型而完全丢弃文档的原始内容, 必然会引起文档的部分信息遗失. 因此, 这种完全抛弃文档创建者的意图而单纯强调用户观点的思想是有缺陷的, 针对该模型存在的不足, 我们提出了一种基于用户查询与文档内容构建文档特征空间的理论模型.

2　基于查询与内容的文档表示模型

单纯利用用户查询与单纯利用文档内容构建文档特征空间的方法分别代表着两种趋向, 即从用户角度与从文档创建者的角度理解文档主题信息的观点, 两种构建文档表示模型的方法各有优缺点. 鉴于此, 本文提出了一种有效地整合两者的优势, 构建新的文档表示模型的思想. 为了简单起见和便于讨论, 在此仅以向量空间模型为例描述新的文档表示模型的构建方法. 基本思路是: 在信息检索初期, 基于文档内容抽取关键词, 以传统的方式构建文档的向量空间. 随着搜索引擎的运行, 用户查询日志将越积越多, 当其所形成的查询样本空间足够大时就提供了有价值的用户隐性反馈信息, 与文档关联的高频查询词将在一定程度上反映出用户的意图. 即从用户的角度来理解, 这些高频查询词较好地描述了对应文档的主题信息. 此时, 就可以在文档表示中逐步引入查询集的信息, 使得文档的特征空间成为查询空间与文档空间信息的整合, 从而提高描述文档主题信息的带权特征词的适应度与可信度, 达到改善检索性能的目的.

随着查询日志的积累, 原始文档集 $\{D\}$可以分为有查询关联的文档和没有查询关联的文档两个部分, 分别用 $\{D_1\}$、$\{D_2\}$表示. 查询集样本空间的积累需要一个过程, 短期内样本空间太小, 不足以准确反映出大量用户的真实意图, 只有当样本空间积累到一定量, 相对稳定时, 才将查询集信息引入到文档表示中来. 为此, 本文提出"查询样本空间稳定性判定"的概念, 定义如下:

设 $\| Q_n \|$和 $\| Q_{n+1} \|$表示相邻的两个周期内查询样本空间的维度, $0< \varphi \leqslant 1$是一个预设的阈值, 如果 $\varphi \leqslant \lambda = \frac{\| Q_n \|}{\| Q_{n+1} \|}$, 则称查询样本空间是稳定的.

对搜索引擎大规模用户日志进行分析, 发现用户的查询具有明显的局部性, 重现率非常高, Silverstein的研究表明, 在被分析的近 5.8亿个用户查询中, 有 4.2亿个用户查询是重复出现的, 冗余度达到 72.41%, 文献[7]对新浪爱问搜索(http://iask.com/)67天的用户查询日志进行分析, 显示在 41 870 667次查询中重复查询有 28 511 271个, 冗余度为 68.09%, 它对日志中每天出现的新查询占总查询的比例进行统计, 表明随着天数的增加, 新查询的比例逐渐减小, 30天后新查询的比例在 28%左右小幅波动, 即查询的重现率(未经过分词处理的 λ的近似值)基本稳定在 72%. 这些数据表明, 查询样本空间确实具有稳定性. 通常当 $0.7 \leqslant \lambda$时, 查询样本空间进入稳定状态. 因此, 可以将 φ预设为 0.7, 然后再根据具体情况予以调整.

为了减少系统的额外负担, 通常将搜索引擎的网络蜘蛛定期更新的周期作为查询样本空间稳定性判定周期. 当 $\lambda < \varphi$时, 查询集太小不足以反映文档的语用信息, 此时文档集 $\{D\}$的特征词权重用传统的向量空间模型方法计算. 随着用户查

询日志的积累, 查询样本越来越大, 当 $\lambda \geqslant \varphi$ 时, 将查询集信息引入到文档表示中来. 此时, 文档集 $\{D_2\}$ 的特征词权重计算方法不变, 文档集 $\{D_1\}$ 的特征词权重由查询空间与文档空间的权重组合而成, 计算方法如下:

用 w_D 表示传统向量空间 D_v (如文档集 $\{D\}$、$\{D_2\}$)中特征词的权重, W_Q 表示查询空间 Q_v 中查询词(至少引发文档被点击一次)的权重, $Q_v \subseteq D_v$, $W_{Q\cap D}$ 表示整合查询空间与文档空间的组合向量空间(文档集 $\{D_1\}$)中特征词的权重. 依据最典型的文档词语权重计算方法 tf* idf[6], 定义如下:

$$W_{D_{ij}} = \frac{tf_{ij}\log(N/n_j + 0.01)}{\sqrt{\sum_{j=1}^{t}(tf_{ij})^2[\log(N/n_j + 0.01)]^2}} \quad (2)$$

其中, tf_{ij} 表示文档 i 中第 j 个词的词频, N 表示文档集中的文档总数, n_j 表示含有第 j 个词的文档数目.

$$w_{Q_{ij}} = \frac{qf_{ij}\log(M/m_j + 0.01)}{\sqrt{\sum_{j=1}^{t}(qf_{ij})^2[\log(M/m_j + 0.01)]^2}} \quad (3)$$

qf_{ij} 表示引发文档 i 被点击的第 j 个查询词的词频, M 表示查询集中查询串的总数, m_j 表示含有第 j 个查询词的查询串的数目.

$$w_{(Q\cap D)_{ij}} = \alpha w_{Q_{ij}} + (1-\alpha)w_{D_{ij}}$$

$$= \alpha\frac{qf_{ij}\log(M/m_j + 0.01)}{\sqrt{\sum_{j=1}^{t}(qf_{ij})^2[\log(M/m_j + 0.01)]^2}} + (1-\alpha)\frac{tf_{ij}\log(N/n_j + 0.01)}{\sqrt{\sum_{j=1}^{t}(tf_{ij})^2[\log(N/n_j + 0.01)]^2}} \quad (4)$$

其中, α 为加权参数. 当 $\lambda < \varphi$ 时, 有 $\alpha = 0$, $(1-\alpha) = 1$, 组合模型退化为传统的向量空间模型; 当 $\varphi \leqslant \lambda \leqslant 1$ 时, 有 $0 < \alpha < 1$, $0 < 1-\alpha < 1$, α 的表达式可以有多种选择, 在特定的环境中需要通过实验来寻找较好的形式.

基于查询与内容的文档表示模型算法简述如下:

输入参数: φ

1) 对文档集 $\{D\}$ 中每个文档 (d_i) 按公式(2)计算文档内容向量空间中特征词的权重;

2) 计算 λ;

如果 $\lambda < \varphi$, 则算法结束;

否则, 对于对文档集 $\{D_1\}$ 中的每个文档 (d_i) 按公式(3)计算查询向量空间中特征词的权重;

按公式(4)计算组合向量空间中特征词的权重;

算法结束;

估算得到文献[2]中第一个模型的 λ 值近似于 0.8 大于 0.7, 符合我们的"查询空间稳定性判定"的假设, 且取得了略优于向量空间模型的性能改善, 一定程度上佐证了我们的新模型的合理性.

3 总结及下一步工作

语法、语义和语用信息三位一体是一个完整的系统. 仅从用户查询或仅从文档内容抽取特征词进行文档表示的方法都存在缺陷和不足, 本文提出的基于查询与内容的文档表示模型, 综合了用户和文档创建者所提供的主题信息, 理论上更加有效. 下一步的工作是在特定的环境中通过实验优化 φ 的取值, 确定 α 的计算公式, 并与其它相关模型进行对比实验, 分析并进一步调整和完善该模型.

参考文献:

[1] 钟义信. 自然语言理解的全信息方法论[J]. 北京邮电大学学报, 2004, 27(4): 1-12.

[2] Poblete B, Baeza-Yates R. Query-Sets Using Implicit Feedback and Query Patterns to Organize Web Documents[C]//Internation WWW Conference. New York: Association for Computing Machinery, 2008, 4: 41-50.

[3] B. P^ossas, Ziviani N, Wagner Meira J, et al. Set-based vector model: An efficient approach for correlation-based ranking[J]. ACM Trans Inf Syst, 2005, 23(4): 397-429.

[4] 张 敏, 宋睿华, 马少平. 基于语义关系查询扩展的文档重构方法[J]. 计算机学报, 2004, 27(10): 1395-1401.

[5] Castellanos M. Hotminer: Discovering hot topics from dirty text[C]//Berry M W. Survey of Text Mining. New York: Springer Science Business Media, Inc., 2003.

[6] Salton G, Buckley C. Term-Weighting approaches in automatic text retrieval[J]. Information Processing and Management, 1988, 24(5): 513-523.

[7] 李亚楠, 王 斌. 一个中文搜索引擎的查询日志分析[J]. 数字图书馆论坛, 2008(7): 1-10.