

文章编号:1001-9081(2008)11-2958-03

基于本体和用户相关反馈的扩展查询研究

王旭阳

(兰州理工大学 计算机与通信学院, 兰州 730050)

(wangxy@lut.cn)

摘要:描述了一种扩展查询(QE)的新方法,这是一种连接用户相关反馈和本体的混合扩展查询技术,有两大贡献:一是连接了用户相关反馈和本体技术,二是采用 FirteX 作为实验平台。与目前广泛应用的基于余弦相似性的扩展查询技术相比,实验结果表明方法平均精度达到 15%,高于基于余弦相似性的扩展查询技术的 13%,并且将平均反馈率提高到了 16%。

关键词:信息检索; 扩展查询; 用户相关反馈; 本体

中图分类号: TP391.1 文献标志码:A

Query expansion based on user relevance feedback and ontology

WANG Xu-yang

(School of Computer and Communication, Lanzhou University of Technology, Lanzhou Gansu 730050, China)

Abstract: A novel method for Query Expansion (QE) was presented in this paper. The proposed method was a hybrid QE technology that combined user relevance feedback and ontology. The proposed method had two significant novelties: it combined user relevance feedback with ontology and used FirteX as the experimental platform. The proposed method was compared with cosine similarity-based QE that was a widely used query expansion technique. The experimental results show that the proposed method outperforms cosine similarity-based QE 15% and 13% in terms of average precision and average recall and has an improvement of 16% in F-measure.

Key words: information retrieval; Query Expansion (QE); user relevance feedback; ontology

0 引言

一个信息检索(Information Retrieval, IR)系统包括数据库、索引和匹配机制。数据库包含一些数据文件,索引连接着用户相关查询和每个文件中相关项目的映射机制。IR 系统的基本目标是从给定的索引数据库中找到一组包含查询者所需信息的文件。在信息检索系统中,扩展查询是目前最热门的研究课题之一。在用户进行查询的过程中,一般的 IR 查询语言需要精确地给出需要查询的信息。也就是说,用户需要精确地表述查询条件。然而,由于大多数用户缺乏查询领域的知识,也许就不能描述清楚他们真正需要的信息。另外,一个普通用户描述的单个查询条件经常会检索到许多相关的文献,但不能找到隐藏在知识或文献内容中的关系。

目前已经有很多关于扩展查询技术的研究工作。文献[1]提出了一种基于模糊规则的扩展查询方法。文献[2]提出了一种用查询日志进行概率统计来扩展查询的方法。文献[3]提出了一种基于词语相似树模型的查询方法。文献[4]提出了一种为扩展查询挖掘查询条件的方法。文献[5]提出了一种通过附加查询条件来挖掘 Web 文件的方法。文献[6]提出了在语义学习对象检索领域的本体扩展查询算法。

本文将介绍一种基于用户相关反馈和本体技的扩展查询方法。运用 FirteX^[7]实验平台、文本检索会议(Text REtrieval Conference, TREC)论文集中的文件作为实验数据进行了检测。为了验证这种方法的性能,我们将它和基于余弦相似度的扩展查询方法^[8]进行比较。实验结果表明这种方法在反馈率和查询精度方面都有很大提高。

1 系统体系结构

基于用户相关反馈和本体的智能信息检索系统的结构如图 1 所示。

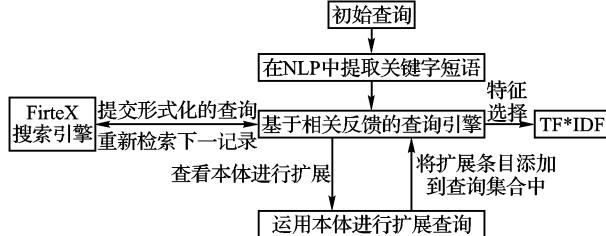


图 1 系统体系结构

该系统使用 FirteX 查询引擎,FirteX 是中国科学研究院计算所开发的,国内第一个开放源码的信息检索实验平台,具有较好的鲁棒性并支持大规模数据,在设计方面,它比广泛应用于信息检索系统中的 Lemur 有更好的性能。

本体在该系统中用于扩展查询。我们首先构造了词语网本体,然后通过词语网去检索符合扩展查询语句和语义规则的相关文件。通过层次遍历词语网找出扩展查询中最适合的入口,但是对于一个给定的条件可能存在大量的查询结果。为了进一步精确查找,我们运用了简单易懂的消除单词技术^[9],它基于词语网短语和关键字之间的相似性。在词语网中,一组具有相同意义的同义词构成了一个同义词集合 synset。这个 synset 自始至终连接在所有相关关系中,如果在所给词组中找不到同义词,则向下遍历 synset 列表继续查找。对于大量的 synset,通过同义词和他们的解释及其描述、包括

收稿日期:2008-05-15;修回日期:2008-07-06。

作者简介:王旭阳(1974-),女,甘肃陇西人,副教授,主要研究方向:智能信息处理、数据库、本体技术。

每个 synset 之间的联系来找到所有非终端单词。然后通过相似度函数 $Sim(S)$ 比较这些单词、术语与关键字列表:

$$Sim(S) = \sum_{i=1}^M \max_{j \in [1, \dots, n_i]} w(p_{ij}) \quad (1)$$

其中 $w(p_{ij})$ 是短语 P_{ij} 在检索文件的同义词集合中出现的频率, 如果不在同义词集合中则为 0。选出同义词集合中相似度最高, 即 Sim 值最大的加入到扩展查询中。

特征选择部件负责从输入文件中选择主要的词语和短语, 我们使用了一种基于关键词提取技术的信息获取系统来选取需要的词语和短语^[10]。另外, 我们运用 TF * IDF 技术^[11]计算文本文件中每个关键词的频率指数, 根据其计算结果过滤不需要的词语。

目前, 用户相关反馈技术也广泛应用于信息检索系统。尽管用户相关反馈相对比较容易实现, 但是让用户从一批文件中找出关键词并加上标记是比较烦琐的, 一般用户最多从摘要列表中选择一个相关文件。但是让检索系统去选择就容易得多, 这正是我们运用搜索引擎的原因。如果初始查询结果非常重要, 我们就考虑一种相关反馈的替换方式, 比如选择匹配等级最大的文件, 这种机制就称为伪相关反馈。

目前有多种方法来处理查询条件的转换, 我们使用了向量机制, 用 $\vec{q} = \alpha\vec{q}_0 + \beta\vec{d}$ 进行转换^[12], 其中: \vec{q} 是查询的向量表示方式, \vec{q}_0 是运用本体的扩展查询表示, \vec{d} 是相关文件, α 和 β 是转换参数。假设文件和查询都是规格化的, 令 $\alpha = \beta = 1$, 则图 1 所描述的方法可概括如下:

1) 用户提供一组初始查询条件, FirteX 系统通过后端索引来检索一个样本文件。

2) 检索文件集合中的每个文件被划分成句子, 然后通过自然语言处理技术从输入文件中选择关键词, 并将它们通过关键字提取系统进一步选取每个句子中的候选码短语, 选中的候选码和关键字长度不超过 3 个连续的词组, 并且不能起止于一个结束符。TF * IDF 对每个候选码短语计算其在文件中出现的频率。最后, 候选码短语根据频率指数划分匹配等级。

3) 运用本体扩展查询技术检索其他与实例匹配的文件。

4) 形式化描述 3) 的查询结果, 参照用户相关反馈并通过搜索引擎再次检索以进一步提高查询结果与初始查询要求的匹配程度。

上述步骤可用图 2 描述。



图 2 系统工作流程

2 算法

系统算法描述如下:

- 1) 通过 FirteX 检索初始请求;
- 2) 运用 TF * IDF 计算主要的关键词在检索文件中出现的频率;
- 3) 运用本体扩展(1)的结果并形式化描述 \vec{q}_0 ;
- 4) 通过用户相关反馈计算 \vec{d} ;
- 5) 计算 $\vec{q} = \alpha\vec{q}_0 + \beta\vec{d}$;
- 6) 带入参数 \vec{q} 在 FirteX 上查询;
- 7) 为信息提取检索 TREC 记录。

该算法对初始查询要求首先在 FirteX 上检索得到一组文件。一旦把数据分开, 则在 TF * IDF 上就能提取出重要的名

词和名词短语, 通过这些名词和名词短语我们再运用本体来扩展初始查询, 也就是形式化 \vec{q} , 然后在 FirteX 中运用 \vec{q} 的值进一步查询并检索记录。

3 实验结果

在实验中, 我们使用第五届 TREC 论文集作为实验数据, 该数据集共 475 MB, 包括 131 896 篇文献, 并把该方法与余弦相似度查询引擎进行比较。余弦相似度模型提出在向量空间进行检索的技术, 用于在文件之间进行相似度计算。在余弦相似的情况下, 两个文件看成是 m 维用户空间的两个向量。它们之间的相似度通过计算两个向量夹角的余弦来衡量。即, 在一个给定的 $m \times n$ 矩阵中, 用 $sim(i, j)$ 函数值来表示 i 向量和 j 向量的相似度:

$$sim(i, j) = \cos(\vec{i}, \vec{j}) = \frac{\vec{i} \cdot \vec{j}}{|\vec{i}| * |\vec{j}|} \quad (2)$$

这里“ \cdot ”表示两个向量的点积。

图 3 和图 4 分别表示了两种方法的查询精度和反馈率。 F 是连接精度和反馈率的衡量参数, 它为信息提取系统提供一个数字化的度量。

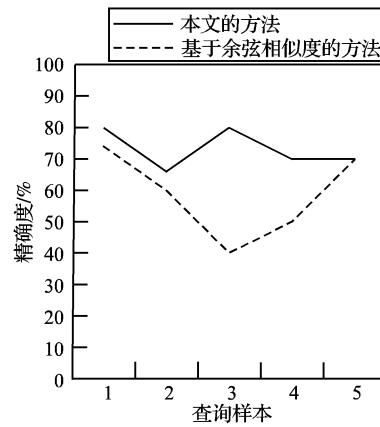


图 3 精确度

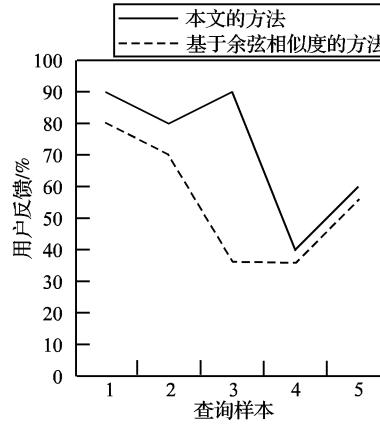


图 4 反馈率

$$F = \frac{2PR}{P+R} \quad (3)$$

这里 P 是精度, R 是反馈率。

系统评估时采用 TREC-10 的测试装置^[13]。测试装置通过搜集用户搜索一个静态文本时提出的查询状态在两个系统中执行的结果进行分析, 其精度、反馈率和 F-参数的值分别如图 3~5 所示。

从图 3~5 可以看出我们的方法比基于余弦相似度的查询方法性能要好, 平均精确度和平均反馈率分别提高了 15% 和 13%。而 F-参数, 我们的方法提高了 16%。在实验过程中, 我们把本体技术和用户相关反馈结合起来进行扩展查询,

避免了查询重构的一些细节,提供了一个用户可控制的过程,用户在反馈中强调相关词语而弱化其他不符合语义的非相关词语,从而提高了查询效率,得到了比较理想的结果。

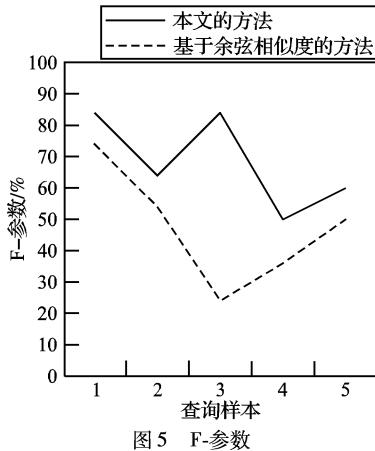


图 5 F-参数

4 结语

本文提出了一种信息检索系统的扩展查询技术。它不同于已有的查询技术,运用 FirteX 系统作为实验平台,并把用户相关反馈和本体技术结合起来进行扩展查询。最后通过一系列的实验表明我们的方法有较好的性能。

参考文献:

- [1] CHANG Y C, CHEN S M, LIAU C J. A new query expansion method based on fuzzy rules[C]// Proceedings of the 7th Joint Conference on AI, Fuzzy system, and Grey system. Taipei, Taiwan, CN: [s. n.], 2003: 335 – 344.
- [2] CUI HANG, WEN JI-RONG, NIE JIAN-YUN, et al. Probabilistic query expansion using query logs[C]// Proceedings of the 11th International Conference on World Wide Web. New York, NY: ACM Press, 2002: 325 – 332.
- [3] JIN QIAN-LI, ZHAO JUN, XU BO. Query expansion based on term similarity tree model[C]// Proceedings of the 2003 International Conference on Natural Language Processing and Knowledge Engineering. Washington, DC: IEEE Computer Society, 2003: 400 – 406.
- [4] LIN H C, WANG L H, CHEN S M. A new query expansion method for document retrieval by mining additional query terms[C]// Proceedings of the 2005 International Conference on Business and Information. Hong Kong, China: [s. n.], 2005: 487 – 496.
- [5] MARTIN-BAUTISTA M J, SANCHES D, CHAMORRO-MARTINEZ J, et al. Mining Web documents to find additional query terms using fuzzy association rules[J]. Fuzzy Sets and Systems, 2004, 148 (1): 85 – 104.
- [6] LEE M C, TSAI K H, WANG T I. A practical ontology query expansion algorithm for semantic-aware learning objects retrieval[J]. Computers and Education, 2007, 50(4): 1240 – 1257.
- [7] 中科院计算所智能软件部. FirteX 全文索引和检索平台[EB/OL]. [2008 - 03 - 01]. <http://www.firtex.org/>.
- [8] SALTON G, BUCKLEY C, FOX E A. Automatic query formulations in information retrieval[J]. Journal of the American Society for Information Science, 1983, 34 (4): 262 – 280.
- [9] SONG M, SONG I Y, HU X H, et al. Integration of association rules and ontologies for semantic query expansion[J]. Data and Knowledge Engineering, 2007, 63(1): 63 – 75.
- [10] SONG M, SONG I Y, HU X H. KPSpotter: A flexible Information gain-based key phrase extraction system[C]// Proceedings of the 5th ACM International Workshop on Web Information and Data Management: WIDM'03. New York: ACM Press, 2003: 50 – 53.
- [11] ZHANG YUN-TAO, GONG LING, WANG YONG-CHENG. An improved TF-IDF approach for text classification[J]. Journal of Zhejiang University Science, 2005, 6A(1): 49 – 55.
- [12] ROCCHIO J J. Relevance feedback in information retrieval [M]// SALTO G. The SMART Retrieval System-Experiments in Automatic Document Processing. Englewood Cliffs, NJ: Prentice-Hall, 1971: 313 – 323.
- [13] TREC. TREC Home Page[EB/OL]. [2008 - 03 - 01]. <http://trec.nist.gov/>.

(上接第 2957 页)

明此算法是可行的,有效的。与基本的 Dijkstra 算法和 A* 算法相比较,节省了大量的时间,效率也得到了极大地提高。

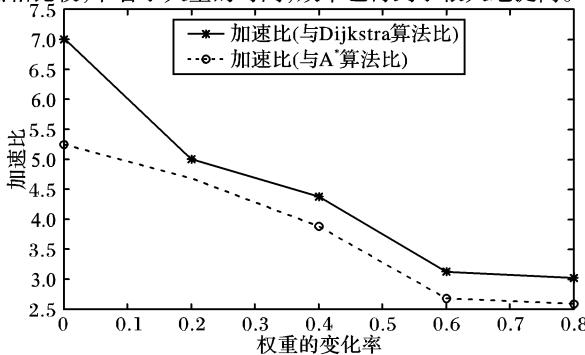


图 4 权重的变化率与加速比的关系

参考文献:

- [1] PAPAGEORIOU M, BLOSSEVILLE J M, HADJ-SALEM H. Macroscopic modeling of traffic flow on the boulevard périphérique in Paris[J]. Transportation Research, 1989, 23(1): 29 – 47.
- [2] LIGHTHILL M J, WHITHAM G B. On kinematic waves II: A theory of traffic flow on long crowded roads[J]. Proceedings of the Royal Society of London, 1955, 229(1178): 281 – 316.
- [3] MICHALOPOULOS P G, YI P, LYRINTZIS A S. Continuum modeling of traffic dynamics for congested freeways[J]. Transportation Research and Application, 2003, 27(4): 315 – 332.
- [4] ZHANG H M. A theory of non-equilibrium traffic flow [J]. Transportation Research-B, 1998, 32(7): 485 – 498.
- [5] ZHANG H M. A non-equilibrium traffic flow devoid of gas-like behavior[J]. Transportation Research-B, 2002, 36(3): 275 – 290.
- [6] 吴正. 低速混合型城市交通的流体力学模型[J]. 力学学报, 1994, 26(2): 149 – 157.
- [7] 戴世强, 冯苏菲, 顾国庆. 交通流动力学: 它的内容方法和意义[J]. 自然杂志, 1997, 19(4): 196 – 201.
- [8] DEAN B C. Shortest path in FIFO time-dependent networks: theory and algorithm [R]. Massachusetts: Massachusetts Institute of Technology, 2004.
- [9] NORDBECK S, RYSTEDT B. Computer cartography shortest route programs[M]. Wseden: The Royal University of Lund, 1969.
- [10] 严寒冰, 刘迎春. 基于 GIS 的城市道路网最短路径算法探讨[J]. 计算机学报, 2002, 23(2): 210 – 215.
- [11] NILSSON N J. Artificial intelligence: A new synthesis[M]. 郑扣根, 译. 北京: 机械工业出版社, 1999.
- [12] ADLER J L. A best neighbor heuristic search for finding minimum paths in transportation networks [C]// Proceedings of the 77th Transportation Research Board Annual Meeting. Washington, DC: National Academy Press, 1998.
- [13] PAPAGEORIOU M. Application of automatic control concept to traffic flow modeling and control[M]. Berlin: Springer-Verlag, 1983.