基于 Ontology的语义查询分析研究

郑世明[1], 任在安[2], 宋自林[1], 邵荣明[3], 戴荣荣[4], 潘明聪[5]

(1. 解放军理工大学 指挥自动化学院, 江苏 南京 210007; 2. 解放军炮兵学院南京分院 侦察教研室, 江苏 南京 211132;
3. 解放军沈阳炮兵学院 海防侦察与射击教研室, 辽宁 沈阳 110162; 4. 解放军 73666部队 炮瞄教研室, 江苏 南京 211132;
5. 南京陆军指挥学院 作战实验中心, 江苏 南京 210045)

[摘要] 目前的搜索引擎普遍存在着查全率和查准率不高的问题, 任何一个简单的查询都可能返回数以万计的检索结果, 而其中只有很少一部分与用户真正的检索要求有关, 对查询的处理是基于本体(Ontology)的语义检索最重要的部分. 针对现有查询分析方法的不足, 提出了一种基于 Ontology的综合词义关系和语义关联分析的查询分析算法, 给出了基于本体映射的语义相似度算法, 通过对用户输入关键字词义特性和本体实例之间语义关联强弱的分析, 提高了用户输入关键字到本体概念映射的完整性和准确率, 保证了用户查询和检索语言在语义上的一致性, 提升了查询的满意度.

[关键词] 本体, 查询分析, 语义, 信息检索

[中图分类号] TP 301 [文献标识码] A [文章编号] 1672-1292(2008)04-0063-05

Research on the Analysis of Semantic Queries Based on Ontology

Zheng Shiming[1], Ren Zaian[2], Song Zilin[1], Shao Rongming[3], Dai Rongrong[4], Pan Mingcong[5]

(1. Institute of Command Automation, PLA University of Science and Technology, Nanjing 211007, China;
2. Reconnaissance Staff Room, Nanjing Artillery Academy of the PLA, Nanjing 211132, China
3. Coast Defense Reconnaissance and Fire Staff Room, Shenyang Artillery Academy of the PLA, Shenyang 110162, China
4. Artillery Collination Staff Room, 73666 Troop of the PLA, Nanjing 211132, China
5. Combat Experimental Center, Nanjing Army Command College of the PLA, Nanjing 210045, China)

Abstract Now search engines generally have the problem of lower recall and precision. Any simple query may return thousands of results for retrieval, but few of results relate with the genuine requirement for the users. It is the most part of semantic retrieval based on ontology to deal with the queries. Aiming at the shortage for existing methods in analysis of query, this paper presents an arithmetic with analysis of query which integrates the acceptation relation and semantic correlative analysis; proposes an arithmetic based on calculation of similarity, analyses the acceptation characteristic of keywords and the acceptation correlative intensity between ontology instances that users input into the system, improves the recall and precision of the mapping from keywords inputted by users to ontology concept, ensures the semantic coherence between query and retrieval language, and promotes the satisfaction for querying.

Key words ontology; analysis of query; semantic; information retrieval

目前, 信息检索技术的算法虽然已经比较完善, 并被广泛应用于各类搜索引擎, 但这些算法很少是基于语义或句法的, 检索时只是从字符的表现形式上进行匹配, 只能检索用户所表达的显性信息, 无法对用户表示隐性信息[1]. 因此, 在检索结果中会出现许多非用户期望的结果, 查准率和查全率较低, 这主要是由于用户、计算机、应用程序三者对用户的查询提问缺乏共同的语义理解. 本文提出了一种基于本体的语义相似度算法和语义查询算法, 通过查询扩展的方法, 提高了用户提问和检索语言在语义上的一致性, 改善了查询效率, 提升了查询的满意度.

1 Ontology相关理论

Ontology最早是一个哲学概念, 从哲学的范畴来说, Ontology是客观存在的一个系统的解释或说明, 是

收稿日期: 2008-06-18

通讯联系人: 郑世明, 博士研究生, 研究方向: 数据挖掘、语义检索. E-mail: zhengshimingnanjing@126.com

客观现实的抽象本质. 在人工智能界最早给出 Ontology定义的是 Neches等人, 他们将 Ontology定义为“给出构成相关领域词汇的基本术语和关系, 以及利用这些术语和关系构成的规定这些词汇外延的规则的定义[2]”.

定义[3] 本体可定义为一个七元组 $O=(C, AC, R, AR, H, I, X)$, 其中 C 是概念的集合; AC 是概念属性的集合; R 是关系的集合; AR 是关系属性的集合; H 表示层次的集合; I 是实例的集合; X 是公理的集合.

1993年, Gruber给出了 Ontology的定义: Ontology是概念模型的明确的规范说明. 后来 Borst在此基础上, 给出了 Ontology的另外一种定义[4]: Ontology是共享概念模型的形式化规范说明. 目前最广为接受的定义是: “本体是对共享概念模型的形式化的明确的描述[5]”, 具有以下 4个方面的特性: 明确性、形式化、共享性、概念模型.

2 查询分析

查询分析是语义信息检索领域内的一个重要内容, 对检索条件不仅包含概念同时也包含语义, 对于用户各种形式的输入, 例如词语或自然语言, 加入特定的背景信息或者语义关系的过程, 从而使计算机能够更好地理解用户输入[6]. 假定我们输入检索词“爱国者”, 我们感兴趣的可能是一些爱国的人群, 或者是“爱国者”导弹, 也或者是“爱国者”移动存储介质, 传统的检索系统不能从用户的输入中获得背景知识, 因而无法进行精确的语义匹配. 在基于 Ontology的语义查询中, 会将检索所需的背景知识如“爱国者”是一移动存储介质传达给查询系统, 从而使得用户和计算机在理解上达到一致.

目前的信息检索系统, 无论是中文还是英文, 大部分都还是基于关键字进行的查询, 本文通过用户输入的关键字, 通过查询分析后进行语义扩展和关联分析, 对扩展后得到的概念的同义词或关联词进行检索, 把用户希望而单凭输入的关键字查询无法检索到的结果返回给用户. 一般有两种扩展方式:

(1) 使用同义词或近义词词典, 例如用户要检索“计算机”, 用“电脑”、“微机”可以表达同样的概念.

(2) 通过本体概念的映射, 分析词之间的语义关联度, 在知识推理机制的基础上实现语义匹配.

第一种方式通常使用包含词与词之间相关信息的资源来进行, 这种方式使用的资源往往经过专家的参与, 能够保证扩展的语义不会有太多损失, 但是这对资源配置的要求比较高, 不一定能够得到相关领域的全面词典. 第二种方式包括使用本体库和知识库, 本体库的可靠性和关联算法是核心, 通过数学的方法自动获得词的共享信息, 在英文信息检索中的应用获得了理想的检索效果[7].

3 本体映射与相似度计算

3.1 本体映射

语义检索中的查询分析方法主要分两类: 关键字匹配和用户浏览. 关键字匹配将用户输入的查询关键字, 通过关键字匹配的方法, 映射到知识库中的类和实例. 文献[8-10]采用了基于关键字匹配的查询分析方法, 在关键字映射到具体的本体概念后, 通过本体关系推导, 发现与用户查询相关的概念. 关键字匹配的特点是简单直观, 但由于对同一概念, 用户输入的查询关键字和本体知识的文字描述往往并不相同, 因而容易产生错误的关键字到本体实例的映射, 并遗漏正确的映射. 与之相对的用户浏览方法, 更多地依赖用户参与来完成查询概念的确定. 文献[11, 12]通过关键字匹配缩小备选查询对象的范围, 最终的查询对象仍由用户确定, 相对于前面的方法, 在映射准确率和用户的时间两个因素间进行了平衡, 取得了一定的效果, 但在解决关键字映射中的同义词、多义词问题和提高映射的准确率方面没有取得实质性的进展.

3.2 本体映射中的语义相似度计算

本文从概念相似度(sim_C)、属性相似度(sim_P)、层次相似度(sim_H)和实例相似度(sim_I)描述经过映射的本体概念之间的语义相似度.

定义 **1** M、N 为本体, M_{C_i} 为本体 M 中的一个概念 C_i, N_{C_j} 为本体 N 中的一个概念 C_j.

定义 **2** $\mathrm{sim}_C(M_{C_i}, N_{C_j})$, $\mathrm{sim}_P(M_{C_i}, N_{C_j})$, $\mathrm{sim}_H(M_{C_i}, N_{C_j})$, $\mathrm{sim}_I(M_{C_i}, N_{C_j})$ 分别为概念相似度、属性相似度、层次相似度和实例相似度.

3.2.1 概念相似度

$$\mathrm{sim_C}(M_{C_i}, N_{C_j}) = \frac{|\,syn(M_{C_i}) \cap syn(N_{C_j})\,|}{|\,syn(M_{C_i})\,| + |\,syn(N_{C_j})\,|},$$

$|\,\mathrm{syn}(M_{C_i})\,|$为本体$M$中概念$C_i$的同义词数; $|\,\mathrm{syn}(N_{C_j})\,|$为本体$N$中概念$C_j$的同义词数; $|\,\mathrm{syn}(M_{C_i}) \cap \mathrm{syn}(N_{C_j})\,|$为本体$M$、$N$中共同包含概念$C_i$、$C_j$的同义词数.

3.2.2 属性相似度

$$\mathrm{sim_P}(M_{C_i}, N_{C_j}) = \frac{\sum_{m=1}^{K}\sum_{n=1}^{L}\mathrm{sim}(M_{C_i}^{(m)}, N_{C_j}^{(n)})}{K \times L},\quad \mathrm{sim}(M_{C_i}^{(m)}, N_{C_j}^{(n)}) = \begin{cases} 1, & \text{属性相同时} \\ 0, & \text{属性不同时} \end{cases},$$

$\mathrm{sim}(M_{C_i}^{(m)}, N_{C_j}^{(n)})$表示$M$本体中概念$C_i$的第$m$个属性与$N$本体中概念$C_j$的第$n$个属性之间的相似度, K、L分别为概念C_i、C_j所包含的属性数.

3.2.3 层次相似度

$$\mathrm{sim_H}(M_{C_i}, N_{C_j}) = \alpha\,\mathrm{sim}(f_{C_i}^{M}, f_{C_j}^{N}) + \beta\,\mathrm{sim}(s_{C_i}^{M}, s_{C_j}^{N}) + \gamma\,\mathrm{sim}(b_{C_i}^{M}, b_{C_j}^{N}),\ \alpha+\beta+\gamma = 1,\ \mathrm{sim}(f_{C_i}^{M}, f_{C_j}^{N}),\ \mathrm{sim}(s_{C_i}^{M}, s_{C_j}^{N}),\ \mathrm{sim}(b_{C_i}^{M}, b_{C_j}^{N})$分别表示概念$C_i$, C_j的父类、子类和兄弟类节点之间的相似度, α、β、γ分别为赋予他们的权重, 由于层次结构中父类、子类和兄弟类节点对语义相似度的影响程度不同, 假设$\alpha \geqslant \beta \geqslant \gamma \geqslant 0$

3.2.4 实例相似度

$$\mathrm{sim_I}(M_{C_i}, N_{C_j}) = \frac{|\,I^{M,N}(M_{C_i})\,| + |\,I^{M,N}(N_{C_j})\,|}{|\,I(M_{C_i})\,| + |\,I(N_{C_j})\,|},\quad |\,I(M_{C_i})\,|$$表示本体M中概念C_i包含的实例数; $|\,I(N_{C_j})\,|$表示本体N中概念C_j包含的实例数; $|\,I^{M,N}(M_{C_i})\,|$表示本体M中属于概念C_i也属于概念C_j的实例数; $|\,I^{M,N}(N_{C_j})\,|$表示本体N中属于概念C_i也属于概念C_j的实例数. 当两个概念的实例个数完全相同时, 其相似度为1; 当两个概念没有相同的实例时, 其相似度为0; 否则相似度为$[0, 1]$上的某个值.

3.2.5 语义相似度

将得到的概念相似度、属性相似度、层次相似度和实例相似度加权求和得到语义相似度Sim: $\mathrm{Sim} = \omega_1\,\mathrm{sim_C}(M_{C_i}, N_{C_j}) + \omega_2\,\mathrm{sim_P}(M_{C_i}, N_{C_j}) + \omega_3\,\mathrm{sim_H}(M_{C_i}, N_{C_j}) + \omega_4\,\mathrm{sim_I}(M_{C_i}, N_{C_j})$, $\omega_1 + \omega_2 + \omega_3 + \omega_4 = 1$, ω_1, ω_2, ω_3, ω_4分别为概念相似度、属性相似度、层次相似度和实例相似度权重, 其大小可由领域专家给定, ω_1, ω_2, ω_3, $\omega_4 \geqslant 0$

4 语义信息检索

4.1 相关定义

定义 **3** 在使用Wordnet进行词义分析时, 每个关键词K_i($1 \leqslant i \leqslant n$)对应于第$j$篇文章的同义词集合$S_i^{(j)} = \{s_{i1}^{(j)}, s_{i2}^{(j)}, \cdots, s_{im}^{(j)}\}$, m为关键词在第j篇文章中的同义词的个数, $|\,S_i^{(j)}\,|$为集合$S_i^{(j)}$中元素的个数.

定义 **4** 将同义词集合中的元素映射到本体库, 得到n个本体集合E_i($1 \leqslant i \leqslant n$), 每个$E_i$是$S_i^{(j)}$中对应本体实例的集合. $dl^{(j)}(k_i)$表示关键词K_i对应于本体中的实例所处的层次, (Dist(k_s, k_t))表示关键词对应本体库中实例(以本体树结构展现)的最短路径, Maxdep$_i$是实例所在本体树的最大深度(max-depth).

$$sr_{ij} = \begin{cases} \dfrac{dl^{(j)}(k_1) + dl^{(j)}(k_i)}{(\mathrm{Dist}((k_1, k_i) \times \mathrm{max}dep_i \times \mathrm{max}(|\,dl^{(j)}(k_1) - dl^{(j)}(k_i)\,|, 1)}, & k_1 \neq k_i \\ sr_{\min}, & k_1 = k_i \end{cases}$$

$r_{ij} = \log\left(\dfrac{N}{|\,S_i^{(j)}\,|} + c\right)$, N为文档总数, C是通过实验得到的调节参数.

4.2 算法实现

根据前面的算法思路, 设计了下面的查询分析算法:

① 接受用户输入, 将关键词K_i保存在集合$K = \{k_1, k_2, \cdots, k_n\}$. K_i为用户输入的第i个关键词, n为关键词个数.

②对信息进行基于 NLU的抽取后[13], 进行词义分析, 得到每个关键词 K_i 的同义词集合 $S_i^{(j)} = \{s_{i1}^{(j)}, s_{i2}^{(j)}, \cdots, s_{im}^{(j)}\}$, $k_i \in S_i^{(j)}$. 设 $s_{im}^{(j)} \in S_i^{(j)}$ 的词义相关度 $r_{ij} = \log\left(\frac{N}{|S_i^{(j)}|} + c\right)$.

③将同义词集合中的元素映射到本体库, 得到 n 个本体集合 $E_i = \{E_{i1}, E_{i2}, \cdots, E_{in}\}$, $(1 \leqslant i \leqslant n)$ 设 $E_{im} \in E_i$ 的词义相关度为 sr_{ij}, sr_{ij} 等于映射到它的文档库中元素的语义相关度.

④计算查询分析的检索关联度 $IR_n = \alpha \sum_{i,j=1}^{i,j=k} sr_{ij} + \beta \sum_{i=1}^{i=k} r_{ij}$, α, β分别为语义相关度和词义相关度的权重系数, $0 \leqslant \alpha \leqslant 1$, $0 \leqslant \beta \leqslant 1$, $\alpha + \beta = 1$.

⑤将集合中元素按检索关联, 即 IR_n 值从小到大排序, IR_n 值越小, 检索关联度越大, 被查询到的可能性越大, sr_{min}, α, β, C 可以根据用户的需求进行局部调整.

5 实验分析

实验通过模拟查询万方数据库中的论文, 验证本文提出的查询分析方法的有效性. 实验环境为: 内存为 512 M, 操作系统为 Windows XP, CPU为 Intel Pentium 4, 2.66 GHz 实验中, 每对查询输入平均对应的实例组合数为 100个, 如果按照简单的关键字映射的方法, 映射的平均正确率仅为 21%, 采用了查询优化方法后, 映射的成功率达到了 80%, 成功率提高了近 3倍. 实验中我们使用了 8对关键字组合查询, 并利用本体实例的说明信息来代替相应的用户查询关键词, 作为优化的查询输入. 经过优化, 平均的准确率从 15%提高到了 52%, 结果如图 1所示.

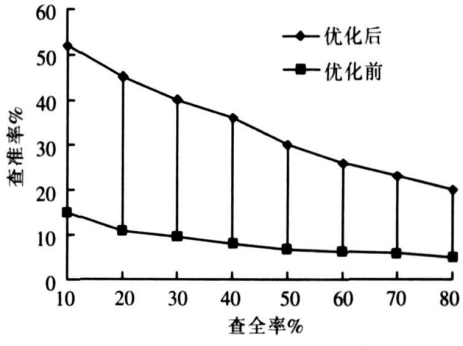

图 1 优化前后查询结果比较

Fig.1 Compare of query result between after and before improving

6 结语

原有的语义查询分析方法大多基于用户浏览或者简单的关键字匹配. 基于用户浏览的方法可以取得较高的分析准确率, 但需要用户付出相当的时间代价. 基于关键字匹配的查询分析简单直观, 但在自然语言含义的不确定性影响下, 往往难以取得让用户满意的分析准确率[14]. 本文针对传统的基于关键字匹配的查询分析方法的不足, 提出了一种综合词义关系和语义关联分析的查询分析方法, 该方法在电子词典和本体知识库支持下, 通过对组成查询关键字的多个词语之间关联的分析, 可以对用户使用关键字表达的查询意图进行有效的推理, 确定查询的具体含义, 提高查询的满意度.

[参考文献] (References)

[1] 冯兰萍, 张继国. 基于本体的中文信息检索模型[J]. 河海大学常州分校学报, 2004, 18(4): 40-41.
Feng Lanping, Zhang Jiguo. Ontology-based Chinese information retrieval model[J]. Journal of Hohai University, 2004, 18(4): 40-41. (in Chinese)

[2] Neches R, Fikes R E, Gruber T R, et al. Enabling technology for knowledge sharing[J]. AI Magazine, 1991, 12(3): 36-56.

[3] 陆建江, 张亚非. 语义网原理与技术[M]. 北京: 科学出版社, 2007: 51.
Lu Jianjiang, Zhang Yafei. The Theory and Technology for Semantic Network[M]. Beijing: Science Publishing Company, 2007: 51. (in Chinese)

[4] Borst W N. Construction of engineering ontologies for knowledge sharing and reuse[D]. Enschede: University of Twente, 1997.

[5] Studer R, Benjamins V R, Fensel D. Knowledge engineering: principles and methods[J]. Data and Knowledge Engineering, 1998, 25(122): 161-197.

[6] 梅翔. 语义检索中若干关键问题的研究[D]. 北京: 北京邮电大学, 2004.
Mei Xiang. Research on semantic search and related technology[D]. Beijing: Beijing University of Posts and Telecommunica-

tions, 2004. (in Chinese)
[7] 王进. 基于本体的语义信息检索研究[D]. 北京: 中国科技大学, 2006.
Wang Jin. Research for semantic information retrieval based on ontology[D]. Beijing: University of Science and Technology of China, 2006. (in Chinese)
[8] Rocha Cristiano, Schwabe Daniel. A hubrid approach for searching in the semantic web[C]// Proceedings of the WWW 2004. New York: ACM Press, 2004: 374-383.
[9] Guba R, McCool R. Semantic search[C]// Proceeding of the WWW 2003. New York: ACM Press, 2003.
[10] 万捷, 滕至阳. 本体论在基于内容信息检索中的应用[J]. 计算机工程, 2003, 29(4): 122-124.
Wan Jie, Teng Zhiyang. Application of ontology in content-based information retrieval[J]. Computer Engineer, 2003, 29(4): 122-124. (in Chinese)
[11] Heflin J, Hendler J. Searching the web with shoe[C]// Artificial Intelligence for Web Search. Menlo park: AAAI Press, 2000: 35-40.
[12] Makela E, Viljanen K, Lindgren P, et al. Semantic yellow page service discovery: The veturi portal[C]// 4th International Semantic Web Conference. Galway, 2005: 65-66.
[13] 李向阳. 基于语义的中文信息抽取研究: [D]. 南京: 解放军理工大学, 2005.
LiXiangyang. Research on Chinese information extract based on semantic[D]. Nanjing: PLA University of Science and Technology, 2005. (in Chinese)
[14] Schumacher M. Security Engineering With Patterns: Toward a Security Core Ontology[M]. Berlin: Springer-Verlag, 2003: 102-103.

[责任编辑: 刘 健]