一种语义搜索中的关联关系排序方法*

文坤梅[1], 李瑞轩[1], 卢正鼎[1], 辜希武[1], 赵艳涛[1]

[1](华中科技大学 计算机科学与技术学院, 武汉 430074)

摘 要: 关联关系搜索是语义搜索中的一种，可发现实体间复杂的关联关系。随着网络上语义资源的迅速增长，实体之间关联关系的个数可能会超过实体本身，因此，在多领域融合的语义搜索中，关联关系排序将是急需解决的问题之一。传统的排序方法主要针对Web资源进行排序，并没有涉及到形式化的语义信息，因此不能用于关联关系排序。针对最常见的路径关联关系，定义了三种影响关联关系排序的重要因素，分别是领域相关度、语义关联长度和关联关系频度，并给出了这些影响因子的权重计算方法，在此基础上提出了一种关联关系结果排序方法。实验结果表明，该方法可优先返回用户真正感兴趣的关联关系，有效地挖掘出实体间有价值的关联关系。

关键词: 关联关系；结果排序；语义搜索；语义Web

中图法分类号: TP301 **文献标识码**: A

传统搜索引擎[1]搜索网络上的各种资源，包括html、word、pdf等多格式文档。随着语义Web技术的发展，产生了大量语义Web资源，xml、本体、Web标注（Web annotation）等语义技术表明了网络资源从数据到知识的转化趋势。斯坦福大学开发的TAP知识库[2]采用RDF描述资源，已涵盖多个不同领域。语义资源相对于传统资源而言，含有丰富的语义信息。传统资源仅描述实体或资源本身，而语义资源则包含资源间的关联关系。随着语义资源的日益丰富，本体中包含了大量关联关系，实体间关联关系的数量已超过实体本身。本体知识库SWETO[3]目前已包含80万个实体和150万个实体间的关联关系，且实体间的关联关系较为复杂。因此，如何从用户角度定义关联关系的重要性，优先返回用户最感兴趣的关联关系，是目前需解决的难点。

1 相关工作

语义搜索将语义Web技术引入搜索引擎，是一个很有研究价值但处于初期阶段的研究课题。近两年来国内外学者采用不同的方法和技术对该课题进行了深入的研究，并得出了不少有益的结论，也建立了相关的原型系统。大多数语义搜索[4, 5]的研究重点集中在如何更好的发现语义实体上，事实上实体之间的关联关系是较实体本身更有价值的语义信息。实体和实体之间通过关联关系紧密结合在一起，孤立的实体本身并不包含语义信息，例如在国家安全、外汇等相关部门，期望通过语义搜索挖掘出实体间有价值的关联关系。因此，相应的研究重点从传统关键字搜索转向对语义资源之间关联关系的搜索。关联关系搜索能够提供一种有效的方法回答“实体X和实体Y之间是否存在某种语义关联”等诸如此类的问题。

关联关系排序的实现基于语义Web与本体技术，同时还依赖于统计学、链接分析、社会网络和词法等相

*Supported by National Natural Science Foundation of China under Grant No. 60873225 (国家自然科学基金); National High Technology Research and Development Program of China (863 Program) No. 2007AA01Z403 (国家高技术研究发展计划(863计划)项目); Open Foundation of State Key Laboratory of Software Engineering under Grant No. SKLSE20080718 (软件工程国家重点实验室开放基金).

作者简介：文坤梅(1979－)，女，博士，讲师，主要研究领域为语义搜索、Web信息管理；李瑞轩(1974－)，男，博士，副教授，主要研究领域为并行计算、语义Web与本体论；卢正鼎(1944－)，男，教授，博导，主要研究领域为数据库系统实现技术、并行计算、异构系统集成；辜希武(1968－)，男，博士，讲师，主要研究领域为Web信息检索、Web服务；赵燕涛(1981－)，男，硕士研究生，主要研究领域为语义Web、关联关系搜索。

关技术。文献[6]提出了一种考虑多因素的关联关系排序方法，该方法考虑的排序因素较多，但实施起来难以保证结果排序的效率。在文献[7，8]中，作者对从大规模的 RDF 元数据中提取语义关联关系。这种方法采用图数据模型的方式表示知识，利用图搜索的方式在图中搜索出路径，即关联关系。SemRank[9]是美国乔治亚州大学研究的一种语义关联关系搜索排序方法，这种方法基于对用户兴趣的可预测性，综合了语义和启发式信息，用户能改变搜索模式得到所需的搜索结果，其排序结果的有效性还有待进一步大规模数据的测试和验证。文献[10]在现代信息检索基础[11]上，利用语义资源的重要性对结果进行排序。国内对关联关系搜索的研究尚处于起步阶段，文献[12]将关联关系用于查询扩展，以提高 P2P 环境下的搜索效果。文献[13]通过关联关系的强弱决定用户意图，以达到提高查准率的目的，但这些相关研究没有考虑到关联关系结果的排序。另外，文献[14，15]也从不同角度对关联关系搜索和结果排序进行了研究。综合来看，关联关系结果排序方法引起了研究者们的广泛关注，但仍未出现一个通用的解决方法，还有待进一步的研究。

本文在已有的研究基础上提出了一种新的关联关系排序方法，方法针对路径关联关系进行排序。和已有的排序方法相比，新的关联关系排序方法定义了影响关联关系搜索结果排序的三种关键因子，并给出其计算方法，试验结果表明排序方法具有较好的可行性，同时保持了较高的查准率。

2 关联关系排序方法

文献[16]定义了三种实体间的关联关系。

（1）路径关联：实体 x、y 之间存在一条属性序列，那么 x、y 是路径关联，表明 x 和 y 的一种直接路径联系。

（2）相交关联：实体 x、y 分别为 ps1、ps2 两属性序列的起点，ps1、ps2 相交于实体 c，那么 x、y 是相交关联，表明从 x 开始的路径 ps1 和 y 开始的路径 ps2 在节点 c 结合。

（3）相似关联：实体 x、y 分别为 ps1、ps2 两属性序列的起点，若 ps1、ps2 中的属性 Pi、Qi 均满足 Pi=Qi 或 Pi∈Qi 或 Qi∈Pi(∈表示子属性关系)，那么 x、y 是相似关联，表明两实体间的相似性。

本文主要考虑第一类关联即路径关联，这也是实体间普遍存在的关联关系。

定义 1 $Q=(O_1,O_n)$ 表示用户查询实体 O_1 和 O_n 之间存在的关联关系，查询结果为关联关系 $R=\{O_1,P_1,O_2,P_2,O_3,.....,O_{n-1},P_{n-1},O_n\}$，其中 O_i 表示实体，P_i 表示实体间存在的关联属性。

好的关联关系排序方法能识别出影响关联关系排序的关键因素。影响关联关系排序的因素包含统计因素、环境因素和人为因素等。本文定义了三种影响关联关系排序的关键因子，分别为领域相关度、语义关联长度和关联关系频度，并给出其计算方法，在此基础上提出了关联关系排序方法 RAR(Ranking Association Relationships)。

2.1.1 2. 1 关联关系影响因子

定义 2 **领域相关度**是指关联关系 R 中所有实体及属性与用户感兴趣领域的相关度，其大小记为 D_R。

用户可能会对某些领域的关联关系更感兴趣，不同的用户其兴趣领域也可能发生变化。设用户的兴趣领域为 D。

属于领域 D 的实体和属性集合为：$Y_i=\{O_i \text{ or } P_i \mid O_i\in R\cap P_i\in R\cap O_i\in D\cap P_i\in D\}$。

不属于领域 D 的实体和属性集合为：$N_i=\{O_i \text{ or } P_i \mid O_i\in R\cap P_i\in R\cap O_i\notin D\cap P_i\notin D\}$。

如学术领域包含与学术相关的所有概念和属性，一般包含实体“教师”、“论文”、“课程”和属性“发表”、“授课”等。

关联关系中属于用户兴趣领域的实体和属性越多，则关联关系的领域相关度越大。关联关系领域相关度

的计算方法如公式（1）所示：

$$D_R = d + (1-d) \times \frac{|Y_i|}{length(R)} \times (1 - \frac{|N_i|}{length(R)}) \quad (1)$$

其中，$length(R)$ 表示关联关系路径长度。d 是为避免 $D_R = 0$ 而设定的调整因子，其大小在 0 和 1 之间，一般取 $0 < d < 0.1$。计算方法表明，领域相关度与关联关系中属于用户兴趣领域 D 的实体和属性个数成正比，与关联关系中不属于领域 D 的实体和属性个数成反比，也即关联关系中属于用户兴趣领域 D 的实体和属性越多，其领域相关度越大。

领域相关度可由用户自行调节与赋值，可划分出若干领域，对不同领域赋予不同权值。相对于一般领域，可赋予用户兴趣领域较高权值。

定义 3 语义关联长度是指关联关系路径长度对关联关系结果排序的影响值，其大小记为 L_R。

对查询结果 $R = \{O_1, P_1, O_2, P_2, O_3, \ldots\ldots, O_{n-1}, P_{n-1}, O_n\}$，其关联路径长度为 n。一般情况下，两关联实体路径长度越短表明其相关度越高，某些情况下则相反。如在国家外汇或安全等部门，用户期望通过复杂的关联关系发现潜在的犯罪嫌疑人或恐怖分子，用户感兴趣的信息可能隐含在较长的关联关系中，较长的关联关系相对于较短的关联关系，应被赋予更高的相关度。

$$L_R = \frac{1}{length(R)} \ or \ L_R = 1 - \frac{1}{length(R)} \quad (2)$$

公式（2）给出了两种计算关联长度的方法，前者表明关联关系的关联路径长度越短，则语义关联长度 L_R 越大，即相关度越高，第二种计算方法则完全相反，对具有较长关联路径的关联关系赋予较大的 L_R。用户可结合实际需求选择合适的语义关联长度计算公式。

定义 4 关联关系频度是指关联关系中所有实体的出入度对关联关系结果排序的影响值，其大小记为 F_R。

类似 PageRank 技术，一个实体具有更大的入度和出度，则表明其具有更高的相关性。如在教育领域，作为“学校”这一概念的两个实体“清华大学”和“长江大学”，在 RDF 图中，“清华大学”具有更大的入度和出度，表明“清华大学”相对于“长江大学”而言，“清华大学”具有更高的相关性。这些具有更高出入度的实体可被看作是更为“重要”的实体，对于包含了“重要”实体的关联关系，在排序时可赋予相对较高的权值。

关联关系频度 F_R 由关联关系中所有实体的出入度大小同时决定。公式（4）给出了计算关联关系频度的方法。

$$F_R = \frac{I_R + C_R}{2} \quad (4)$$

其中，I_R 是关联关系 R 的入度，C_R 是关联关系 R 的出度，若关联关系 R 中包含 n 个实体，I_R 和 C_R 的计算方法如下：

关联关系 R 的入度 I_R 由关联关系中所有实体的入度同时决定，可得：

关联关系 R 的入度 $I_R=\frac{1}{length(R)}\sum_{i=1}^{n}\frac{I_i}{Max(I)}$，其中 I_i 是实体 O_i 的入度，$Max(I)$ 是 n 个实体中的最大入度数。

关联关系 R 的出度 C_R 由关联关系中所有实体的出度同时决定，可得：

关联关系 R 的出度 $C_R=\frac{1}{length(R)}\sum_{i=1}^{n}\frac{C_i}{Max(C)}$，其中 C_i 是实体 O_i 的出度，$Max(C)$ 是 n 个实体中的最大出度数。

因此公式（4）等价于：

$$F_R=\frac{1}{2length(R)}\sum_{i=1}^{n}(\frac{I_i}{Max(I)}+\frac{C_i}{Max(C)}) \tag{5}$$

其中 I_i 是实体 O_i 的入度，$Max(I)$ 是 n 个实体中的最大入度数，C_i 是实体 O_i 的出度，$Max(C)$ 是 n 个实体中的最大出度数。

2.1.2 2. 2 关联关系排序方法

关联关系的最终排序结果，需综合考虑各种因素。以上文提出的三种关键影响因子为基础，提出一种关联关系排序方法 RAR。

对查询 $Q=(O_1,O_n)$，其查询结果为关联关系集：

$R_i=\{O_1,P_{1i},O_{2i},P_{2i},O_{3i},\ldots\ldots,O_{(n-1)i},P_{(n-1)i},O_n\},i=1,2,\ldots,m$。

关联关系排序方法 RAR 对关联关系集中的每个关联关系计算其重要性权值 V_R，计算方法如公式(6)所示。

$$V_R=k_1\times D_R+k_2\times L_R+k_3\times F_R \tag{6}$$

其中，$k_1+k_2+k_3=1$，不同用户其排序要求也不相同，用户可根据实际需要对其赋值，D_R 表示领域相关度大小，L_R 表示语义关联长度大小，F_R 表示关联关系频度大小。上述公式（1）、公式（2）和公式（5）分别给出了 D_R、L_R、F_R 的计算方法。在实际应用中，如果用户需要寻找某些特定领域的关联关系，则可对 k_1 赋予较高的权重；如果用户认为路径长度对关联关系的影响较大，则可对 k_2 赋予较高的权重；如果用户需要寻找包含较多重要对象的关联关系，则可对 k_3 赋予较高的权重。

公式（6）表明，关联关系的重要性权值与领域相关度、关联关系长度和关联关系频度的大小成正比。根据公式（6）计算出的关联关系重要性权值 V_R，对查询所得到的关联关系结果进行排序，V_R 越大，表明该关联关系越重要，将被优先返回给用户。

3 方法实现与性能分析

3.1.1 3．1 搜索实现

对关联关系排序之前，需搜索出实体间存在的所有关联关系。利用深度优先的图搜索算法实现关联关系搜索，采用邻接矩阵作为图的存储结构。寻找邻接点所需的时间为 $O(n^2)$，其中 n 为图中的顶点数。可通过改变数据结构及优化搜索算法来改进搜索效果，在此不作探讨。

选择规模不同的四个测试本体，对应的载入及搜索时间如表 1 所示。当本体中实例个数大于 1000 时，对不同实体间的关联关系查询，搜索时间差异过大，此时无法计算平均搜索时间。从图中可得搜索时间与本体大小成正比。

Table 1 Ontology loading and searching time

表 1 本体载入及搜索时间

本体编号	本体文件名	本体大小	实例个数	平均载入时间	平均搜索时间
1	Semrank.owl	8kb	10	31ms	85ms
2	Animal.owl	30kb	75	47ms	250ms
3	Idc_onto.owl	102kb	98	141ms	1813ms
4	Apex_Portal_0.99.owl	821KB	1128	516ms	~~~

3.1.2 3．2 排序方法实现

根据上文提出的关联关系排序方法 RAR，实现了语义搜索中关联关系排序的原型系统。

载入本体时需将本体中所有的类和域名空间提取出来，以便于用户选择领域。用户输入待查询的两实体，点击“搜索”按钮，文本框中将显示排序后的所有关联关系。系统采用 Jena 实现本体解析，界面如图 1 所示。

长关联优先（Long Association）：若选中，则优先返回路径较长的关联关系，否则优先返回较短的关联关系。

长度权值（Length weight）、领域权值（Context weight）和关联关系频度权值（Node In&Out weight）：图中用椭圆标识出来，这三个值分别表示语义关联长度，领域相关度，关联关系频度在排序算法中所占的权重，其取值范围都是 0 至 1 之间，且三者之和等于 1。

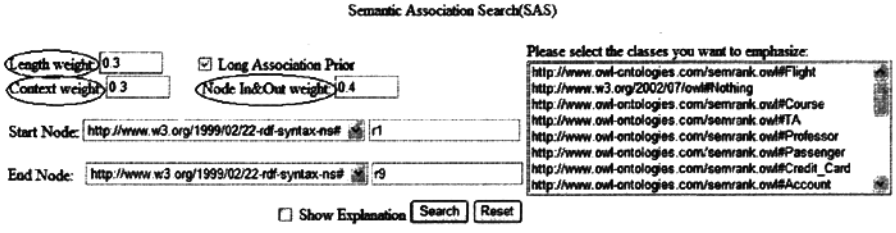

Fig.1 Search interface of association relationships

图 1 关联关系搜索界面

起始点域名（Start Node）、终点域名（End Node）：载入本体时从本体中提取出来，用于限制起始点和终点的范围，区分不同本体中具有相同的起始名或终点名。

领域列表框（Emphasize classes）：载入本体时领域列表框显示本体中所有的类名，用户选择零个或多个类限定领域范围，搜索结果中包含较多所选类的路径优先返回。

排序值细节（Show Explanation）：在每条关联路径的后面显示详细的排序算法计算过程，关联关系的重要性权值利用上述排序公式（6）计算。

图 2 表示搜索实体 r1 和 r9 之间的关联关系，结果显示搜索耗时 390 毫秒，搜索到 6 条相关纪录，即 r1

和 r9 之间存在 6 条关联路径，且根据 RAR 方法计算出的关联关系重要性权值进行了排序。

Time: 390 ms Existing 6 semantic association relationship.

http://www.owl-ontologies.com/Semrank.owl#r1 and http://www.owl-ontologies.com/Semrank.owl#r9 have following association relationship!

1 r1 **depositsIuto** r8 **AcctHolder** r6 **ownsStockIn** r9
Total Value: 0.40803033

2. r1 **adviseeof** r6 **ownsStockIn** r9
Total Value: 0.3976767

3. r1 **ownsStockIn** r5 **electedLeader** r6 **ownsStockIn** r9
Total Value: 0.39666668

4 r1 **bitsFor** r2 **forFlight** r3 **paidBy** r4 **AcctHolder** r5 **electedLeader** r6 **ownsStockIn** r9
Total Value: 0.38497838

5. r1 **audits** r7 **taughtBy** r6 **ownsStockIn** r9
Total Value: 0.3830303

Result pages: 1 2 Next page

Fig.2 Search results of association relationships between r1 and r9

图 2 实体 r1 和 r9 之间的关联关系搜索结果

图 3 表示搜索“文坤梅”和“孙小林”之间的关联关系，结果显示搜索到 5002 条相关记录。

Fig.3 Search results of association relationships between “文坤梅” and “孙小林”

图 3 “文坤梅”和“孙小林”之间的关联关系搜索结果

3.1.3 3. 3 性能评测与分析

目前对语义搜索还没有统一的评测标准。可借鉴传统的搜索评测方法，将实际值与期望值进行对比以评测关联关系排序方法。

（1）单本体测试

以表 1 中本体 Semrank.owl 为测试用例，该本体涉及多个领域，且实体间具有较复杂的关联关系。

定制五种典型组合进行测试，在每个测试查询中，强调两种排序影响因子（即赋予它们较高的排序权重值）。表 2 列出了查询操作及其对应的测试目的。五个不同的用户对排序方法进行测试，考虑到不同的用户对

结果排序时具有一定的主观性，因此将所有用户的平均排序结果看作排序理想值。

Table 2 Query operations and test purposes

表 2 查询操作及测试目的

序号	查询操作	测试目的
1	查询“Passenger”和“Organization”实体间的关联关系，强调短路径和包含“运输”类（如“Ticket”和“Flight”等）的关联关系。	测试排序方法获取短路径、兴趣领域相关的关联关系的能力。
2	查询“Customer”类型的两实体间的关联关系，强调长路径和包含“组织”类（如“Organization”）的关联关系。	测试排序方法获取长路径、兴趣领域相关的关联关系的能力。
3	查询“Customer”类型的两实体间的关联关系，强调长路径和包含重要节点的关联关系。	测试排序方法获取长路径、包含重要节点的关联关系的能力。
4	查询“Customer”类型和“Account”类型的实体间的关联关系，强调包含“组织”类（如“Organization”）和重要节点的关联关系。	测试排序方法获取兴趣领域相关、包含重要节点的关联关系的能力。
5	查询个“Customer”类型的两实体间的关联关系，综合考虑三种影响因子。	根据用户需求确定排序标准后的排序结果。

测试结果如图 4 所示，图中给出了系统排序结果和用户排序结果的交集，它显示了系统与人为排序结果的一致性，其中“理想排序”表示一种理想情况，即系统和人为排序结果完全吻合。如图可知，用户排序的结果与系统排序结果较为接近，有部分排序结果是直接匹配的。

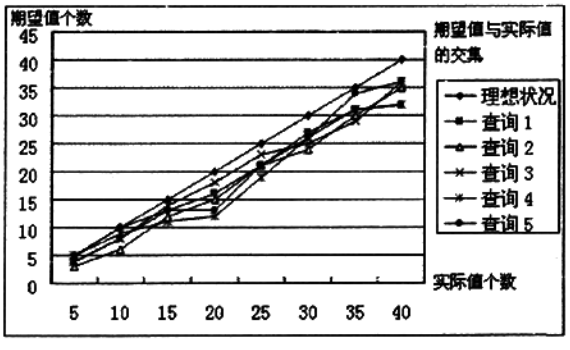

Fig.4 Test results of RAR

图 4 RAR 测试结果图

根据查准率计算公式 $P=\frac{|R_a|}{|R|}$，其中 $|R_a|$ 表示检索出的相关文献的个数，$|R|$ 表示相关文献总数，可得查询 1 对应的平均查准率为 83.6%；查询 2 对应的平均查准率为 76.5%；查询 3 对应的平均查准率为 86.4%；查询 4 对应的平均查准率为 81.8%；查询 5 对应的平均查准率为 88.7%；总的平均查准率为 83.4%。测试结果表明，查询平均查准率在 80%以上，用户查询结果与期待结果的偏差在可接受范围之内。

（2）多本体测试

以表 1 中的多个本体为测试用例，用户随机选择关联关系查询，评测结果取其平均值。每个关联查询的结果选取前 20 个。给出八种影响因子排序组合，表 3 列出了查询说明及其测试目的。对每一个可能产生的查询，在测试之前通过手工形式给出理想的关联关系排序结果，理想结果的手工制定者和实验测试者不是同一

个人，但都是计算机领域的学生，对于关联关系搜索比较熟悉。对每一个查询，首先返回搜索得到的 20 个未排序的关联关系，理想结果制定者们基于对关联关系搜索和该查询的理解，给出他们认为正确的关联关搜索结果排序结果。考虑到不同的用户对结果排序时具有一定的主观性，将所有用户的平均排序结果看作排序理想值。

Table 3 Query description and test purposes

表 3 查询说明及测试目的

查询序号	查询说明	测试目的
Q1	查询两实体之间的关系，其所属类别相同	屏蔽领域相关度和关联关系频度，测试查询算法能否优先获取关联长度较长的结果
Q2	查询两实体之间的关系，其所属类别相同	屏蔽领域相关度和关联关系频度，测试查询算法能否优先获取关联长度较短的结果
Q3	查询两实体之间的关系，设定特定“学术”为用户感兴趣领域	屏蔽关联路径长度和关联关系频度，测试查询算法能否优先获取用户感兴趣领域的结果
Q4	查询两实体之间的关系，连接两者的某一条路径中包含有出入度频繁的结点	屏蔽关联路径长度和领域相关度，测试查询算法能否优先获取包含更多“热门”结点（关联关系频度较大）的结果。
Q5	查询两实体之间的关系，设定路径长短和兴趣领域	测试关联路径长度和领域相关度的组合
Q6	查询两实体之间的关系，设定路径长短偏好且关联路径中包含有出入度频繁的结点	测试关联路径长度和关联关系频度的组合
Q7	查询两实体之间的关系，设定兴趣领域且关联路径中包含有出入度频繁的结点	测试领域相关度和关联关系频度的组合
Q8	查询两实体之间的关系，设定路径长短和兴趣领域且关联路径中包含有出入度频繁的结点	测试关联路径长度、领域相关度和关联关系频度三个影响因素的组合

其中，Q1 到 Q4 属单因素测试，而 Q5 到 Q8 属多因素测试。为测试排序方法的有效性，给出了期望值与实际结果的对比图，图中显示了查询语义关联结果与用户期望结果的交集数。该图表明了期望值与实际值之间的关联关系。理想查询阐明的是一种理想关系，说明期望值与实际查询结果完全吻合，下图中 I 表示理想值。图 5-12 表示查询 Q1 到 Q8 的测试效果。

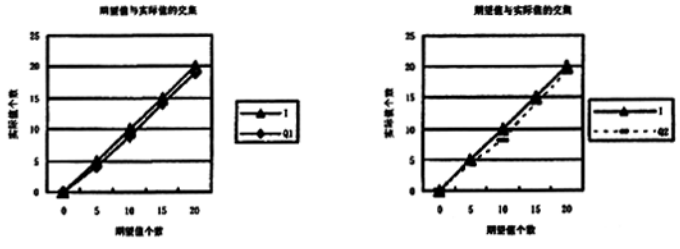

Fig.5 testing results of Q1

图 5 查询 Q1 的排序交集测试结果

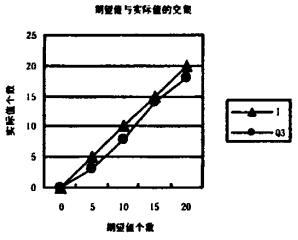

Fig.6 testing results of Q2

图 6 查询 Q2 的排序交集测试结果

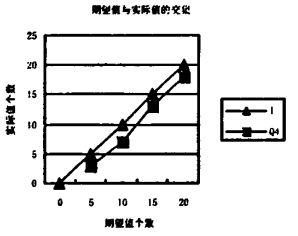

Fig.7 testing results of Q3

图 7 查询 Q3 的排序交集测试结果

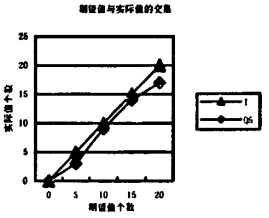

Fig.8 testing results of Q4

图 8 查询 Q4 的排序交集测试结果

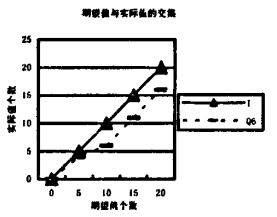

Fig.9 testing results of Q5

图 9 查询 Q5 的排序交集测试结果

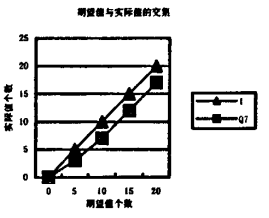

Fig.10 testing results of Q6

图 10 查询 Q6 的排序交集测试结果

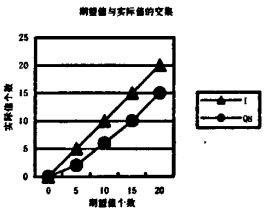

Fig.11 testing results of Q7

图 11 查询 Q7 的排序交集测试结果

Fig.12 testing results of Q8

图 12 查询 Q8 的排序交集测试结果

测试结果表明，单因素测试排序方法时，搜索结果与理想结果吻合度较高，多因素测试时，因不同的用户对结果有不同的标准，因此存在一定的偏差，搜索效果有所降低。根据平均查准率计算公式 $\overline{P}=\frac{1}{8}\sum_{i=1}^{8}P_i$，其中 P_i 表示查询 Qi 对应的查准率，通过计算可得关联关系排序方法的平均查准率在 80%以上，用户查询结果与理想状况的偏差在可接受范围之内，该排序方法能较好的将用户希望搜索的关联关系优先返回给用户。尽管不同用户的排序标准存在分歧，但测试结果表明了排序方法的可行性，且该方法的灵活性可满足不同用户的多种偏好，能让用户获得满意的搜索结果。

4 结束语

关联关系搜索发现实体之间的复杂关系，随着语义资源的迅速增长，实体之间的关联关系可能会超过实体本身，因此关联关系搜索结果排序是语义搜索中需要迫切解决的问题。针对最普遍的路径关联关系，本文提出了三种影响排序的关键因素，分别是领域相关度，关联关系长度及关联关系频度，定义了这三种关键影响因子的计算方法，以此为基础提出了一种关联关系排序方法 RAR，试验结果表明该方法能较好地将用户需要的关联关系优先返回。如何用更大规模的数据集进一步测试方法的有效性，并将方法扩展到其他更多类型的关联关系搜索中，是下一步研究目标。

References:

[70] Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, et al. Searching the web. ACM Transaction on Internet Technology, 2001, 1(1): 2-43.

[71] Guha R, McCool R. TAP: A Semantic Web Test-bad. Journal of Web Semantics, 2003,1(1)

[72] http://lsdis.cs.uga.edu/projects/semdis/sweto/.

[73] Guha R, McCool R, Miller E. Semantic search. Proceeding of the 12th International World Wide Web Conference (WWW 2003). Budapest, Hungary, May 2003:700-709.

[74] Kunmei Wen, Zhengding, Xiaolin Sun, Ruixuan Li. A Semantic Search Conceptual Model and Application in Security Access Control. V4185 LNCS, ASWC 2006,366-376

[75] Boanerges Aleman-Meza, Christian Halaschek-Wiener, I. Budak Arpinar, et al. Ranking Complex Relationships on the Semantic Web. IEEE Computing, 2005,9(3):37-44.

[76] Anyanwu, K., Sheth, A. The ρ operator: Discovering and Ranking Semantic Associations on the Semantic Web, ACM SIGMOD Record, v.31 n.4, December 2002: 690-699

[77] Halaschek, C., Aleman-Meza, B., Arpinar, B., Sheth, A.Discovering and Ranking Semantic Associations over a Large RDF Metabase. VLDB 2004 demo paper: 254-264

[78] Kemafor Anyanwu, Angela Maduko, Amit Sheth. SemRank: Ranking Complex Relationship Search Results on the Semantic Web International World Wide Web ConferenceWWW 2005: 117-127

[79] Bhuvan Bamba, Sougata Mukherjea. Utilizing Resource Importance for Ranking Semantic Web Query Results. SWDB2004,185-198

[80] BAeza-Yates and Ribeiro-Neto. Modern Information Retrieval. Addison Wesley 1999

[81] Zhang Q, Zhang X, Liu JR, Sun Y, Wen XZ, Liu Z. Query expansion and its search algorithm in hybrid peer-to-peer networks. *Journal of Software*, 2006,17(4):782−793. http://www.jos.org.cn/1000-9825/17/782.htm

[82] Wang Guiling, Jiang Jinlei, Shi Meilin. Context query and association discovery for collaborative environment. *Journal of Sountheast University (English Edition)*, 2006,22(3):338-342

[83] Stojanovic N, Studer R and Stojanovic L. An approach for the ranking of query results in the semantic web. In Proc. of ISWC 2003

[84] Aleman-Meza B, Halaschek C. Arpinar I.B, Sheth A. Context-Aware Semantic Association Ranking. First Intl. Workshop on Semantic Web and DBs, Berlin, Germany 2003

[85] Anyanwu K, Wheth A. ρ-queries: enabling querying for semantic associations on the Semantic Web. In proceedings of the 12th International Conference on World Wide Web, 2003:690-699.

附中文参考文献:

[12] 张焘,张霞,刘积仁,孙雨,文学志,刘铮. 混合 P2P 环境下有效的查询扩展及其搜索算法. 软件学报. 2006,17(4):782−793. http://www.jos.org.cn/1000-9825/17/782.htm

[13] 王桂玲,姜进磊,史美林. 协作环境中的上下文查询和关联发现. 东南大学学报(英文版). 2006,22(3):338-342

A Ranking Method for Association Relationships in Semantic Search *

WEN Kun-Mei, LI Rui-Xuan, LU Zheng-Ding, Gu Xi-Wu

(School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China)

Corresponding author: Phn: +86-27-62164762, Fax: +86-27-87544285, E-mail: knwen@hust.edu.cn

Abstract: Searching association relationships is a kind of semantic search. It can find out complicated relationships between entities. As the quick growth of Semantic web, the number of association relationships is possibly greater than the number of entities themselves. So how to rank association relationships is becoming the new important question of semantic search. Not considering semantic information, traditional ranking methods are designed for traditional Web resources. So these methods are not feasible for searching association relationships. Aiming at the common path association relationships, three critical influence factors are defined. They are domain related degree, semantic association length and semantic association frequency. The methods of computing three factors are designed. Based on these work, the method used to rank semantic association relationships is proposed. The experiment results show that the method can return the most useful semantic association relationships to users and efficiently dig out the valuable association relationships between entities.

Key words: Association relationships; Result ranking; Semantic search; Semantic web