

基于词义消歧的语义查询扩展研究

罗俊丽¹, 李慧娜¹, 路 凯²

(1 许昌学院 计算机科学与技术学院, 河南 许昌 461000)

(2 许昌电气职业学院, 河南 许昌 461000)

摘 要: 为了解决传统查询扩展时查准率低下的问题, 基于词义消歧技术提出一种综合扩展语义树和词频共现率的语义查询扩展方法. 针对查询词歧义所带来的查询主题漂移现象, 利用 WordNet 知识源及其领域信息进行查询词义消歧, 进而根据 WordNet 的层次结构生成扩展语义树, 产生候选扩展词, 并根据待扩展词与用户查询的整体最大相关性原则最终确定扩展词及其权重, 使得扩展词能够充分表达用户查询请求, 提高查询匹配准确率. 实验表明, 该方法在保证查全率的同时获得了较高的查准率.

关键词: 词义消歧; 信息检索; 语义扩展; WordNet

中图分类号: TP301

文献标识码: A

文章编号: 1000-7180(2012)01-0071-05

Semantic Query Expansion Research Based on Word Sense Disambiguation

LUO Jun-li¹, LI Hui-na¹, LU Kai²

(College of Computer Science and Technology, Xuchang University, Xuchang 461000, China)

(Xuchang Electrical Vocational College, Xuchang 461000, China)

Abstract: In order to solve the problem of lower precision with the traditional way of query expansion, a semantic query expansion method based on word sense disambiguation is proposed, which integrate the semantic expansion tree and the co-occurrence of words. For the query topic shift aroused from word sense ambiguous, the query word sense is confirmed by disambiguation based on WordNet knowledge and their domains information, and then the candidate expansion terms are generated through the expansion semantic tree which is generated from hierarchy of WordNet, the final expansion terms are determined according to the Maximum principle of the whole correlation of candidate expansion terms and all query words. User request is fully expressed by these expansion words. Lastly, experimental results indicate that our method has a higher precision while ensuring the recall.

Key words: word sense disambiguation; information retrieval; semantic expansion; WordNet

1 引言

在目前基于关键字的文本检索中, 用户使用查询词的平均长度为 2.4 至 2.7 个词, 短小的查询无法精确表达用户的查询意图, 此外, 由于自然语言中存在的多义性, 用户所使用的查询词与文本集中的索引词不匹配, 这些问题阻碍了信息检索性能的提高^[1].

查询扩展技术根据用户的查询, 选取与原始查询相关联的词, 组成更长、更准确的查询, 在一定程度上弥补用户查询信息的不足. 作为提高检索性能的一种有效方法, 查询扩展技术已成为信息检索中的重要研究课题^[2]. 传统的查询扩展方式盲目选择扩展词, 没有考虑原始查询词歧义性, 这样容易因为原始查询词的歧义造成扩展查询“主题偏移”. 此外, 在选择扩展词时, 往往基于原始查询中的单个查询

项来选择,没有充分表达用户的查询意图.针对这些问题,本文提出了一种在词义消歧的基础上进行语义查询扩展的方法(SQEWSD).在选取扩展词时,即考虑了其单个原始查询项的语义相关性,还考虑了扩展词与所有原始查询项在初检相关文档中的整体语义相关性.从而使所选取的扩展词更能接近用户的查询意图,提高信息检索的准确率.

2 基于词义消歧的语义查询扩展

该算法的基本思想是:首先利用 WordNet^[3]知识库及其领域信息确定查询词的词义,以此避免查询主题偏移;然后根据确定的查询词义,从 WordNet 的层次结构中提取出扩展语义树,并选取候选扩展词;最后,采用平均互信息理论,评估每个候选扩展词与原查询在初检文档中的整体相关性,并最终确定扩展词.这样就使得扩展词与用户查询主题具有较强的语义相关性,提高信息检索准确率.

2.1 查询词义消歧

(1) 构造语义关联图 G

由于用户查询短小,在文中采用初次检索结果中的相关文档作为消歧上下文.以查询词及其上下文中包含的所有实词在 WordNet 中对应的所有词义作为图节点,以 WordNet 中定义的各种语义关联关系作为词义节点间的无向边,将语义关联的强度作为边的权重.当两个节点间的语义关联有多种时,把强度最大的关联作为边权重,此加权无向图即为消歧的语义关联图 G.参照 WordNet 中节点间语义关系的强度,对 G 中无向边的权重进行人工定义.

(2) 基于改进的 PageRank^[4] 算法进行词义消歧

原始的 PageRank 算法应用在一个边权重均等的有向图上,要想将该算法应用在无向的语义关联图 G 上,而且连接边的权重各不相同,就必须对原始的 PageRank 算法进行改进才能够运用到词义消歧中.

在改进的 PageRank 算法中,节点在关联图 G 中的重要性值不仅与投票数、投票节点的重要性值有关,还与投票的强弱(边权)有关,这三者共同决定了节点的 PageRank 值.设 $G=(V, E)$ 是一个具有节点集 V 和边集 E 的无向图,则节点 V_i 的 PageRank 值定义如下:

$$Pr(V_i) = (1-d) + d \times \sum_{j \in link(V_i)} Pr(V_j)$$

$$\times \left[\frac{w_{ji}}{\sum_{k \in link(V_j)} w_{jk}} \right] \quad (1)$$

其中, w_{ij} 表示由节点 V_i 到节点 V_j 的边权重, $link(V_i)$ 表示与节点 V_i 相连接的节点集, d 是一个制动因子,取值范围为 $(0, 1)$.

对于任意分配的图节点的初始 PageRank 值,利用改进的 PageRank 算法迭代计算各节点的 PageRank 值,直到图中节点的 PageRank 值收敛为止.此时,在查询词的所有词义节点中,PageRank 值最高的节点代表的词义作为消歧后的确切词义:

$$S_{GR}(w, C) = \{S_i \mid \forall S_k \in w, Pr(S_i) \geq Pr(S_k)\} \quad (2)$$

若一个节点具有多个最高 PageRank 值,将保留多个词义作为目标词的词义,一般来说,这些词义是非常接近的.

(3) 词汇的领域相关性

研究表明,词汇的领域属性有助于提高消歧精度.因此, SQEWSD 方法通过判断查询词词义与上下文词汇的词义间的领域相关性来进一步提高消歧的性能.在计算领域相关性时,借鉴了 Magnini 等人对领域属性的归纳思想,用 43 个领域标签概括 WordNet Domains 中的所有领域属性,以此提高领域匹配的概率.同时考虑词频信息 f 对消歧的影响,词义 S_i 和 S_j 之间领域相关性的计算方法如下:

$$DomSIM(S_i, S_j) = \begin{cases} 0, & \text{if } Dom(S_i) \cap Dom(S_j) = \phi \\ \frac{1}{f_i} \times \frac{1}{f_j}, & \text{if } Dom(S_i) \cap Dom(S_j) \neq \phi \\ & \wedge Dom(S_i) \neq Factotum \\ 0.1 \times \frac{1}{f_i} \times \frac{1}{f_j}, & \text{if } Dom(S_i) = Dom(S_j) \\ & = Factotum \end{cases} \quad (3)$$

词义 S_i 和上下文词汇 C_k 之间的领域相关性定义为 S_i 和 C_k 的所有词义的领域相关性的极大值,即:

$$DR(S_i, C_k) = \max_{S'_k \in w_k} (DomSIM(S_i, S'_k)) \quad (4)$$

对于目标词 w 的每个词义 S_i ,依次计算其与所有上下文词汇的领域相关性,然后求和得到词义 S_i 与上下文 C 的整体领域相关性如下:

$$DomR(S_i, C) = \sum_{k=1}^{|C|} DR(S_i, C_k) \quad (5)$$

其中, $|C|$ 为上下文 C 中包含的开放词汇个数.

最后,选择整体领域相关性最大的词义作为目标词的最终词义:

$$S_{Dom}(w, C) = \{S_i \mid \forall S_k \in w, DomR(S_i, C) \geq DomR(S_k, C)\} \quad (6)$$

(4) 查询词义确定

将目标词的所有词义在两种方法中计算得到的重要性值分别进行降序排序,记录其排序(Rank 值)依次为 $1, 2, \dots, n$. 然后,将得到的 Rank 值相乘,并考虑了词频属性对消歧的影响,计算如下:

$$Rank(S_i) = Rank_g(S_i) \times Rank_d(S_i) \times Freq(S_i) \quad (7)$$

其中, $Rank_d(S_i)$ 和 $Rank_g(S_i)$ 分别表示词义 S_i 采用基于领域相关性和基于图论进行消歧得到的 Rank 值, $Freq(S_i)$ 是 S_i 的词频排序值.

则排序最前的词义为查询词的消歧词义:

$$S_{Rank}(w, C) = \{S_i \mid \forall S_k \in w, Rank(S_i) \leq Rank(S_k)\} \quad (8)$$

2.2 扩展词的选取

当查询词的词义确定下来之后,考虑利用 WordNet 本体中的层次结构关系对用户查询进行基于语义的扩展,具体的扩展可分为以下几个步骤.

(1) 扩展语义树的生成及候选扩展词的确定

首先以各个查询项词义为中心,找到其在 WordNet 中的公共祖先节点.把公共祖先节点、查询词义节点及其下层的子节点,连同它们的树状结构一起裁剪出来,称为扩展语义树,它能够最大限度的表达用户查询意图.

在扩展语义树中,离查询项词义节点越远的节点,其概念语义与查询项语义相关程度越弱.设查询项词义节点的权重为 2,则其子概念节点的权重为 2^{-L} , L 为子概念节点与其对应的查询项节点的距离,即相对于其查询项节点的层数.公共祖先节点到查询项词义节点的路径节点概念权重为 2^{-2L} (L 意义同上).公共祖先节点和查询项词义节点的其它父节点的权重等于其到与之相连的各个查询项词义节点的权重之和,该权值表明公共祖先节点、查询项词义的其他父节点和与之相连的所有查询项词义节点的关联程度.

给出如下实例来了解扩展语义树及候选扩展词权重,设原始查询向量为 $Q = \{q_1, q_2, q_3, q_4\}$, 则 Q 的扩展语义树及节点权重如图 1 所示.

(2) 扩展词的选择及其权重计算

语义查询扩展的目标是找到在语义上接近整个查询 Q 的扩展词,而不仅仅是接近 Q 中的某个独立的词.因此,必须评估每个候选扩展词与全部查询项的语义关联度.关于大规模语料的研究表明,当两个

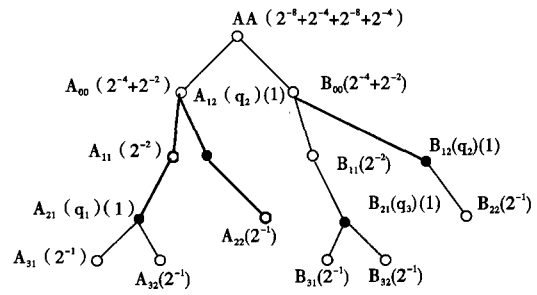


图1 查询语义树 QST 及其节点权重

词经常共同出现时,则认为这两个词在语义上是相互关联的,而且共现频率越高,其语义关联程度越强.把每个候选扩展词与原查询的全部查询项,在前 n 篇初检文档集中的共现程度作为评定候选扩展词与原查询整体语义关联的标准.

Rosenfeld 提出用平均互信息 (Average Mutual Information, AMI) 评估词的关联度^[5],即:

$$AMI(x, y \mid S) = p(x, y) \log \frac{p(y \mid x)}{p(y)} + p(x, \bar{y}) \log \frac{p(\bar{y} \mid x)}{p(\bar{y})} + p(\bar{x}, y) \log \frac{p(y \mid \bar{x})}{p(y)} + p(\bar{x}, \bar{y}) \log \frac{p(\bar{y} \mid \bar{x})}{p(\bar{y})} \quad (9)$$

其中, $p(x, y) = \frac{c(x, y)}{c(x)} + \frac{c(x, y)}{c(y)}$, $p(x) = \frac{c(x)}{\sum_x c(x)}$, $p(x \mid y) = \frac{p(x, y)}{p(y)}$. $c(x, y)$ 是词 x 和 y

在训练集 s 的文档中同一句子内共现的次数, $c(x)$ 是词 x 出现在训练集 s 中的次数.但是即使在一个窗口单元中,随着词间距离的增大,词与词之间的相关程度也会指数递减.在公式(9)中增加一个因子 $e^{Space(x, y)}$ 来表达这种特性,将词 x 和 y 的相似度定义为:

$$SIM(x, y \mid S) = \frac{AMI(x, y)}{e^{Space(x, y)}} \quad (10)$$

其中, $Space(x, y)$ 是指在所有窗口单元中,词 x 与 y 之间的平均相距字数.

假设原查询 Q 中的每个查询项之间相互独立,可以根据候选扩展词 w 与 Q 中每个查询项 q ,在初检出的前 n 篇文档集 s 中的平均互信息之和的对数,来度量 w 和 Q 的语义关联度 (Cohesion Degree),其定义为:

$$Cohd(w, Q \mid S) = \log \left(\sum_{q \in Q} (idf(w \mid C) \times idf(q \mid C) \times SIM(w, q \mid S) + 1) \right) \quad (11)$$

其中, $idf(x|C) = \log \frac{N}{df(x|C)+1}$, $df(x|C)$ 表示语料集 C 中出现词 x 的文档数, N 为语料集 C 中文档总数. 加入 $idf(x|C)$ 是为了避免语料中有些词出现频度可能很高, 但其重要性不是很大(最坏的情况是停用词)也会获得很高的 $Cohd$ 值. 加 1 是为了避免出现为 0 的情况.

综合考虑候选扩展词与查询 Q 的全部查询项的语义关联度 $Cohd(w, Q|S)$ 和它在扩展语义树中的权重, 定义候选扩展词 w 的权重为:

$$Weight(w) = QSTWeight(w) \times Cohd(w, Q|S) \quad (12)$$

其中, $QSTWeight(w)$ 是候选扩展词 w 在扩展语义树(QST)中的权值. 设定阈值 θ , 满足 $Weight(w) > \theta$ 的词即为查询扩展词, 至此查询扩展完成.

3 实验与分析

(1) 实验数据

采用 TREC Novelty 数据集来测试文中所提出的查询扩展的有效性, 依据给定的查询主题, 经过查询扩展后, 返回相关的句子. 数据包括: TREC-2004 提供的 50 个查询主题(编号 N51-N100), 给定的文档数目有 1808 个, 每个主题给出的文档集合包括 25 篇相关文档, 而其它文档则为不相关的噪音数据集. TREC 主题部分包括 <num> 编号、<title> 标题、<totype> 主题类型、<desc> 简述、<narr> 详细描述. 其中 <totype> 有两种: 事件类(Event)与评论类(opinion). 测试时采用 WordNet 提供的接口 WordNet, Net 来构造概念树.

(2) 评测标准

文中引入了 TREC 规定的标准评估方法——平均准确率(MAP)来进行相关的评测. MAP 的定义为:

$$AP = \frac{1}{R} \sum_{i=1}^R \frac{i}{Doc(i)}, \quad MAP = \frac{1}{Q} \sum_{j=1}^Q AP_j$$

其中, R 表示单个查询检索结果中与查询句相关的文档数, $Doc(i)$ 表示第 i 篇相关文档的排名值, AP 表示单个查询的平均准确率, Q 表示查询的总数, MAP 表示所有查询问题的平均准确率的加权平均值.

同时, 实验中还采用信息检索的传统评测方法——查全率和查准率来评估扩展词的选取阈值及初检结果中的相关文档数.

(3) 实验设计及结果

查询扩展是对初始查询的补充和优化, 其扩展词数量会影响到检索的精度, 为此, 实验对选择扩展词的阈值 θ 进行了测试, 结果如图 2 所示.

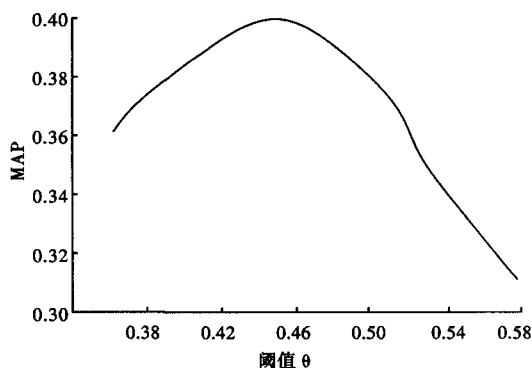


图2 扩展词数对 MAP 的影响

由图 2 可知, 当 θ 设定为 0.46 时查询性能最好, 当 $\theta < 0.46$ 时查询性能下降较快, 这是因为随着 θ 的减小, 加入权重低的扩展词不但不会优化初始查询, 反而会加入噪音词降低查询精确度.

在实现 SQEWSO 方法时, 引入候选扩展词与全部查询项的平均互信息来选取扩展词, 因此初检结果中的相关文档数也会对检索性能造成影响, 其测试结果如图 3 所示.

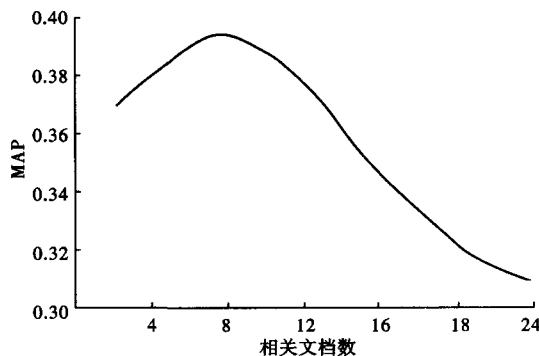


图3 相关文档数对 MAP 的影响

由图 3 可知, 取前 8 篇文档为相关文档时, 可达到最佳的检索性能. 当选取的相关文档数过多的话, 会造成准确率下降, 这是因为选取过多的文档时, 与查询主题不相关的文档比率也相对会越高, 增大了扩展不相关词语的概率.

在前两个测试的基础上, 选取 $\theta=0.46$, 相关文档数 $n=8$, 分别比较了 SQEWSO 方法与基于语义扩展、基于统计扩展、未扩展方法的查全率和 MAP, 结果如图 4 所示.

从比较结果可以看出: 文中提出的 SQEWSO 方法在提高查全率的同时也具有较高的准确率, 其

次分别为基于语义的方法、基于统计的方法和未扩展方法。

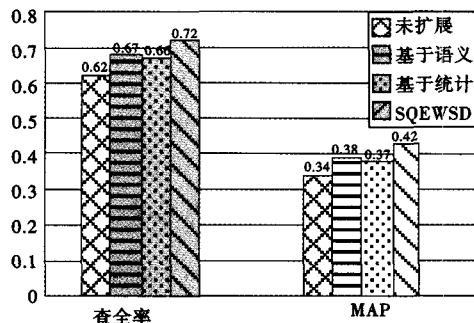


图4 实验结果比较

4 结束语

本文提出了一种基于词义消歧的语义查询扩展方法,在确定查询词词义的基础上,从概念语义层次上实现查询扩展。结合语义查询树与局部反馈技术筛选扩展词,使得扩展词不仅与用户查询中的单个查询项具有较强语义关联,还与整个查询具有较好的语义关联度,保证了扩展词在语义上更加接近用户查询语义,从而有效提高了查询扩展的准确率。下一步将选取不同的测试集来检查该算法的稳定性。

参考文献:

- [1] Furnas G W, Landauer T K, Gomez L M, et al. The vocabulary problem in human-system communication [J]. Communication of ACM, 1987, 30(11): 964-971.
- [2] 王瑞琴. 基于语义处理技术的信息检索模型研究[D]. 杭州:浙江大学, 2009.
- [3] Fellbaum C. WordNet: an electronic lexical database [M]. Cambridge, Massachusetts, UK: MIT Press, 1998.
- [4] Rada M, Paul T, Elizabeth F. PageRank on semantic networks with application to word sense disambiguation[C]//Proceedings of the 20th international conference on Computational Linguistics. USA: Jackson, 2004:1126-1132.
- [5] 蔡平,王志强,傅向华. 基于语义的跨媒体信息检索技术研究[J]. 微电子学与计算机, 2010, 27(3): 102-105.

作者简介:

- 罗俊丽 女, (1985—), 硕士, 助教. 研究方向为信息检索。
 李慧娜 女, (1980—), 硕士, 讲师. 研究方向为信息处理和信息安全。
 路凯 男, (1979—), 硕士, 助教. 研究方向为信息安全。

(上接第70页)

- [2] Bushnell M L and Agrawal V D. Essentials of electronic testing for digital, memory, and mixed-Signal VLSI Circuits [M]. USA: Kluwer Academic Publishers, 2000: 13-14.
- [3] 廖薇, 许春冬, 刘锦高. 基于神经网络的模拟电路故障诊断研究 [J]. 微电子学与计算机, 2010, 27(5): 125-127.
- [4] 黄海, 李华伟. 组合电路中考虑串扰的布尔过程波形模拟 [J]. 微计算机信息, 2010, 26(11-2): 173-174.
- [5] Banerjee S, Chowdhury D R, Bhattacharya B B. An efficient scan tree design for compact test pattern set [J]. Proceedings of the 19th International Conference on VLSI Design, 2006: 3-7.
- [6] Shuyan Jiang, Yongle Xie, Dajin Yu. Experiment and

Simulation of Transistor Level Fault Model of IDDT Test [D]. Chendu: Chendu School of Automation Engineering University of Science and Technology of Chendu. 2009.

- [7] 于云华, 石寅. 数字集成电路故障测试策略和技术的研究进展 [J]. 电路与系统学报, 2006, 9(3): 83-91.
- [8] 马超, 陈西宏, 邓均明, 等. 模拟电路故障诊断的信息融合新方法 [J]. 微电子学与计算机, 2011, 28(4): 10-12.

作者简介:

- 蔡炼 男, (1982—), 博士研究生, 讲师. 研究方向为集成电路测试。