一种基于语义单元的查询扩展方法*)

李 莉 高庆狮
(北京科技大学信息工程学院 北京 100083)

摘 要 查询扩展技术通过向初始查询请求中加入相似或者相关的词，来减少查询请求与相关文献在表达上的不匹配现象，改善检索性能。本文利用语义单元的语义表达能力和语义单元之间的关系，将与初始查询具有密切语义关系的查询词或短语加入到初始查询请求中，更加全面地表示了用户的查询意愿。算法的时间复杂度为 $O(L)$，只与搜索请求的长度 L 有关，与语义单元表示库的规模无关，这对实时性要求较高的搜索引擎来讲是很实用的。

关键词 信息检索，查询扩展，搜索引擎，语义单元

A Query Expansion Method Based on Semantic Element

LI Li GAO Qing-Shi
(Information Engineering School, Beijing University of Science and Technology, Beijing 100083)

Abstract Query Expansion technology can reduce word mismatch between query and related documents, improve retrieval result through adding similar or related terms to original query. In the algorithm proposed in this paper, terms or phrases which have closely related sense are added to the original query and express users' query intention more precisely. This algorithm costs $O(L)$ time which is independent of the SER-Base, and this is very practical for highly real-time search engine.

Keywords Information retrieval, Query expansion, Search engine, Semantic element

1 引言

在网络信息检索系统中，语言中大量存在的同义、多义、上下义等语义关系使得语言的表达方式多种多样。而用户提交的初始查询请求是用户搜索意愿的某一种表达方式，这种特定的表达形式使得原本具有语义联系，但是以其他表达形式表示的相关网页无法被搜索引擎搜索到，导致相关文献与查询请求之间的词有不匹配现象。查询扩展技术通过在初始查询请求中增加具有相似意义或统计相关的词或词组，组成更为准确的扩展查询请求，来减少查询请求与相关文献在表达上的不匹配现象，改善检索性能。

目前常用的查询扩展方法可大致分成三类：基于语义知识辞典的方法、全局分析方法和局部分析方法。语义知识辞典扩展方法借助于 WordNet、HowNet 等语义知识辞典，选择与初始查询词存在一定语义关联性的词来进行扩展，这种方法对语义知识辞典有较高的要求。由于通用的语义知识辞典很难建立，语义知识辞典扩展方法大多在某一具体领域的知识辞典中应用；全局分析方法是对全部文档中的词或词组进行相关分析，计算每对词或词组间的关联程度，根据预先计算的词间相关关系将与查询用词关联程度最高的词或词组加入原查询以生成新的查询。常见的全局分析方法包括词聚类[1]、潜在语义标引(LSI)[2]、相似性词典[3]、基于概念的查询扩展方法[4]等。局部分析利用初次检索得到的与原查询最相关的 N 篇文章作为扩展用词的来源，最有代表性的局部分析方法是相关反馈方法(Relevance Feedback)以及在其基础上发展起来的伪相关反馈方法(Pseudo Relevance Feedback)[5]。近年来，查询扩展技术出现了与其他自然语言处理技术相结合的趋势，例如文[6]将潜在语义分析方法与文献聚类结合起来，文[7]将相关反馈与文本分类技术结合起来，文[8]利用用户的查询日志对词与词之间关系进行分析，自动选择与原查询高度相关的词与词组来进行查询扩展；随着对本体研究的关注，基于本体的检索技术也得到了较快的发展，基于本体的查询扩展技术[9]利用本体所推导出的语义信息对用户的查询请求进行扩展，得到不错的检索效果。

语义单元是在句子中表达一定意思的单元。本文利用语义单元的语义表达能力和语义单元之间的语义关系对用户的查询请求进行扩展，经实验验证，对改善搜索引擎的搜索性能有良好的作用。

2 语义单元

2.1 语义单元的基本概念

语义单元与语义语言的基本概念和形式定义在文[10]中有详细的介绍，为了保持行文的连贯，此处只做简要介绍。任何一种具体的自然语言(如英语，汉语，…)中一个句子的“语义”，称之为“句义”。在句义中表达一个“意思”的单元被称为“语义单元”。句义由若干个语义单元交错构成(这一点不同与通常的“句子由词线性排列而成”)。任何一个具体的自然语言中的一个表达某一个意思(即语义单元)的单元称为该语义单元在该具体自然语言中的“语义单元表示”。一个具体的自然语言 I 的句子是一个句义在该具体的自然语言 I 中的“句义表示”。从这个角度看，句子(句义表示)是由(带变量和不带变量的)语义单元表示构成，两个自然语言(I,J)之间的翻译可

*)本课题得到国家 863 高技术研究发展计划项目基金(2006AA012140)和国家自然科学基金项目(60573014)的资助。李 莉 博士研究生，主要研究方向为自然语言处理、机器翻译。

以看作为两种表示之间的转换。不同的具体自然语言之间的句子之所以能彼此翻译，使用不同的具体自然语言的人们之所以能彼此交流，就是因为不同的具体自然语言之间有对应于相同语义的句子，或者可以建立表达该语义的一组句子。

例如，英语的句子："This coat belongs to you"，是句义"这件外套的归属权是你"在英文中的表示，句义在中文中的表示是"这件外套是我的"。该句义可以写成"$B_{elong\ to}$（T_{his}（coat），you）"，它由"$B_{elong\ to}(X,Y)$"、"$T_{his}(N)$"、"coat"、"you"这四个不可弃语义单元组成，其中"$B_{elong\ to}(X,Y)$"和"$T_{his}(N)$"中的参数 X、Y 和 N 均表示事物义。在计算机系统内部可以表示为表 1。

表 1　语义单元与语义单元表示库

SE	Parameter	SER-ch	SER-en	Type
1	2,N,N_2	N 是 N_2 的	N belong to N_2	J
2	1,N	这 N	this N	N
3	0	外套	coat	N
4	0	你	You	$N_{人}$

语义单元有带参数（如"this N"）和不带参数（如"coat"）两种，对应三种语义单元表示：实量（例如"coat"）、纯虚量（如"A N"，其中 A 是形容词，N 为名词）、实量虚量混合（例如"this N"）。

语义单元能清楚地表达一个语言单位（如句子、短语等）的语义结构，如语言单位所描述的事件、该事件的主体和客体、事件所发生的时间或者位置等语义信息。通过求解句子的语义结构，可以很容易地理解句子的深层语义，进而对句子进行处理（如翻译、对话等）。

2.2　语义表达式的求解

考察已有的语义单元表示库，语义单元表示中实量虚量所有的可能排列可以规范成 4 种基本形式（S、SX、XS、XSX）及其之间的连接，如表 2 所示。其中 S 表示实量串，X 表示虚量串，"|"表示匹配开始，"&"表示两个基本形式之间的连接。除了纯虚量串 X 以外，所有的匹配比较总是从实量开始，然后是实量前的虚量，最后才是实量后的虚量。

表 2　语义单元表示由 4 种基本形式（S，SX，XS，XSX）和连接（$）构成

S-X 排列类型	实现（|表示开始比较）
S	|S
SX	|SX
SXS	|SX$|S
SXSX	|SX$|SX
SXS…XS	|SX$\|S…X$|S
SXS…XSX	|SX$\|S…X$|SX

S-X 排列类型	实现（|表示开始比较）
X	|X
XS	X|S
XSX	X|SX
XSXS	X|SX$|S
XSXSX	X|SX$|SX
XSXSX…S	X|SX$\|SX…$|S
XSXSX…SX	X|SX$\|SX…$|SX

为了快速求解语义表达式，对语义单元表示库进行处理。例如，实量以 know 开始的语义单元表示"know（*，about·N*，N·（$about N_2*），how·to·V*，J*，$N_{人}$·$well*，$N_{人}$·（$from $N_{人_2}$））"可以分解为一个主表示"know（*，about·N*，N·$1，how·to·$V$*，$J$*，$N_{人}$·$well*，$N_{人}$·$2）"和一个子表示集"{（$1：about N_2*），（$2：from $N_{人_2}$）}"，其中"$"标识语义单元表示中第二个以及其后的实量串。主表示是在用语义单元表示序号代替"$"之后的部分，子表示集是由各语义单元表示以及"$"之后的部分构成的。

把语义单元表示库中含两个及以上实量串的语义单元表示分解为主表示和子表示集，用主表示修改语义单元表示库，把子表示集并入库中，经过排序等处理程序，形成新的语义单元表示库；把新的语义单元表示库转换为树形，称为语义单元表示树。上文中分解后形成的语义单元表示可以转化为树形：

know（*，about·N*，N·$1，how·to·$V$*，$J$*，$N_{人}$·（$well*，$2））；

about（N_2*4）；　　←$1

from（$N_{人_2}$*）；　　←$2

对于给定的语言单位 S，求解其语义表达式的快速算法如下：

while（S 非空）

Begin

从 S 中相续地取一个字（w）；

取以该字（w）为开始的最多一棵语义单元表示树；

对已经取出的所有语义单元表示树，根据字（w）进行剪枝，把与（w）不同的枝剪掉；

对类型流进行剪枝处理；

End；

从剪枝后没有被剪掉的语义单元表示中求语义表达式 SSE，并输出.

假设语言单位 S 的长度为 L，语义单元库中语义单元表示的数目为 N，语义单元表示的平均长度为 l，m 是语义单元表示树结点的平均分支数目。在语义表达式求解算法中，取 L 棵语义单元树的时间为 $O(Llm)$；根据 S 中第 $x+1$ 个字 W 对已取出的数目为 $x(x<l)$ 的语义单元表示树进行剪枝，每个节点平均比较 m 次，总剪枝时间是 $(0+1+2+\cdots+l+l+\cdots+l)m=O(Llm)$；因此总的计算时间为 $O(Llm)=O(L)$。也就是说，不论语义单元表示库的规模有多大，算法的时间是固定不变的，这对于实时性要求比较高的搜索引擎来说，是非常适合的。

3　语义单元之间的语义关系

语言中存在着大量的同义、多义、上下义等关系，表现在语义单元表示库中，语义单元之间也存在各种语义关系，如同义、反义、上下位、部分与整体等关系。当然，这些关系仅仅存在同类型语义单元之间，例如事物义语义单元之间、动作义语义单元之间等等。根据信息检索的特征，大部分检索涉及的都是事物、事件、动作等类型，而不涉及修饰功能的词类型，所以重点分析如下几种关系：

（1）事物义语义单元之间的同义关系

根据替代原理，定义语义单元之间的同义关系为：如果两个语义单元具有相同的参数，并且在句中互相替换后不改变句子的意义，则称这两个语义单元同义。例如：电脑与计算机、网络与网路等。

（2）事物义语义单元之间的上下位关系

指一个语义单元相对于另一个语义单元的有限等级，其中一个语义单元是另一个语义单元的次类。上下位关系产生层次语义结构，使得下位词可以继承上位词的所有特征/性质。例如，car 和 vehicle 之间就是上下位关系，其中 car 是 vehicle 的下位词，vehicle 是 car 的上位词。这样我们在搜索 vehicle 的时候，和 car 相关的内容也应该包含在结果内。

(3)动作义语义单元之间的近义关系

动作义语义单元之间很少有真正的同义关系，但是通常存在相同感情色彩或者相同态度等的近义关系，例如反对“台独”、打倒“台独”与批判“台独”等。

为了有效地表示这些语义关系，在语义单元库中增加若干个字段：在事物义语义单元中增加同义字段和下位词字段，在动作义语义单元中增加近义字段，形成扩展语义单元，示例如表 3 所示。

表 3 扩展语义单元与扩展语义单元表示库示例

语义单元	参数	汉语表示	英语表示	类型	同/近义词集	下位词集
$1	0	作品	works	N	($2,…)	($m, $m+1,…)
$2	0	作品	production	N	($1,…)	($m, $m+1,…)
…	…	…	…	…	…	…
$m	0	小说	novel	N	($n*,…)	($n,…)
$m+1	0	音乐	music	N	Φ	($m*,…)
…	…	…	…	…	…	…
$n	0	科幻小说	sci-fi	N	Φ	($x*,…)
…	…	…	…	…	…	…

4 基于语义单元的查询扩展方法

4.1 原理

用户提交给搜索引擎的查询请求通常都是若干个关键词或短语，但是许多情况下几个简单的关键词不足以准确全面地表达出用户的搜索意愿、专业领域等，这对缺乏知识理解能力的搜索引擎的检索性能有比较严重的影响：返回的搜索结果有成千上万页面，这些页面或多或少都与查询相关联，但是与搜索请求的相关度高低不等；不同主题、不同领域的页面混杂在一起，使得用户查找到真正需要的信息很困难。

从搜索引擎的工作流程来看，用户提交的初始查询通常要被搜索引擎进行分词等预处理，预处理过程也是引起语义丢失的原因之一。例如：“中国反对侵略”被切分为“中国”、“反对”和“侵略”三个关键词，只要网页中出现了这三个关键词，不管网页的真实含义是“中国反对侵略”，还是“反对侵略中国”，或是“反对中国侵略”，该网页都会被列入搜索结果中。因此，简单地用若干个关键词来表示用户的查询请求是不准确的，需要能够更精确表达其深层语义的工具。

另一方面，语言中大量存在的同义、上下义等语义关系使得用户的搜索意愿有多种不同的表达形式，用户向搜索引擎提交的只是其中的一种表达形式，而以其它形式表示的页面也应该包含在结果中。例如，电脑与计算机、网络与网路是同义关系，当用户搜索“电脑”时，搜索引擎应该能够根据同义关系将与“计算机”相关的网页搜索出来；又如，搜索某人的作品，搜索引擎也应该可以根据“作品”与“著作”、“电影”、“电视剧”、“音乐”之间的上下位关系，把与作品相关的著作、电影、电视剧、音乐等网页也搜索出来。也就是说，从语义上对搜索任务进行一定的补充和扩展，使得搜索结果更加全面，提高查全率。

鉴于上文中讨论的语义单元的语义表达能力，考虑用语义单元表达用户的初始查询，以提高查准率。用户的查询请求具有一定的语义结构，这种语义结构经过分词等预处理程序后会被破坏，因此我们不对用户的查询请求进行分词等预处理，而是经过语义分析，得到语义表达式之后，借助于语义单元之间的关系进行查询扩展，向初始查询中加入相关的词或短语，以提高查全率。

4.2 算法设计

Begin

接收用户的初始查询字符串 Str；

对 Str 进行语义分析，得到 Str 的语义表达式 SE；

如果 SE 为零参数事物义语义单元，取 SE 的同义词集 $T(Str)$，依次用 $T(Str)$ 中的每一个同义词 $T(Str)_i$ 进行替换，得到扩展搜索请求 $NEStr = \sum T(Str)_i$；

如果 SE 中含有事物义参数 N_k，取 N_k 的同义词集 $T(N_k)$ 和下义词集 $X(N_k)$，依次对 N_k 进行替换展开，得到扩展搜索请求 $NEStr = \sum T(N_k)_i + \sum X(N_k)_j$；

如果 SE 为动作义语义单元 V_k，取 V_k 的近义词集 $J(V_k)$，依次用 $J(V_k)$ 中的每一个近义词 $J(V_k)_i$ 对 V_k 进行替换，得到扩展搜索请求 $VEStr = \sum J(V_k)_i$；

Str 的扩展请求 $EStr = Str + NEStr + VEStr$；

调用现有搜索引擎的接口对 $EStr$ 进行搜索，得到搜索结果集 RS；

End

注：$T(Str)_i$、$X(N_k)_j$、$J(V_k)_i$ 均为不可拆分的扩展搜索串，也就是说使用这些搜索串在现有的搜索引擎中进行搜索的时候不能再进行分词处理，如在 Baidu 和 Google 中要加上双引号；而“+”运算符在搜索引擎中表示“或”运算，即各个加和项的结果集的合集。

设用户初始查询字符串的长度为 L，求解其语义表达式的时间为 $O(L)$；对于现有的语义单元表示库，平均的扩展查询次数是常数 T，那么整个扩展查询的时间复杂度为 $O(L)+T=O(L)$。也就是说，扩展查询算法的时间只与初始查询请求字符串的长度 L 有关，这正是搜索引擎的实时性所要求的。

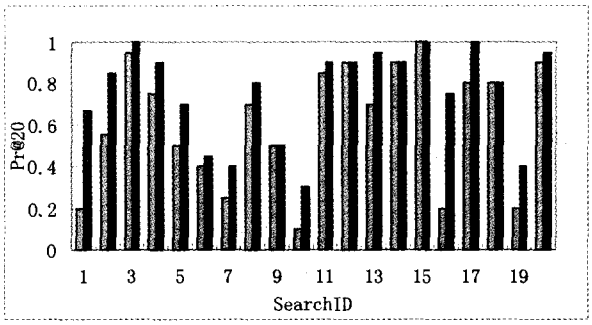
5 实验结果与分析

本实验基于 Google 搜索引擎，将用户初始查询请求和扩展查询请求分别提交给 Google，不涉及搜索引擎实现的细节。

实验选用燕穹数据[9] Web 日志数据(编号 YQ-QUERY-LOG-2004-02)(即天网用户查询日志)作为测试数据，数据的提供方为北京大学互联网信息研究所。作者从日志数据中去除掉暴力、色情等方面的查询数据和重复数据后，抽取 20 个查询请求，得到 20 个测试数据。基于事先已经建好的扩展语

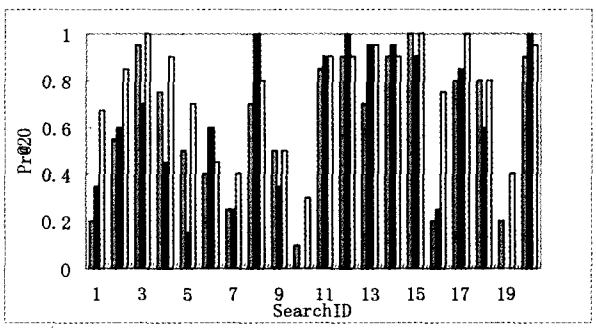
义单元表示树库，实现扩展查询，并将扩展查询请求在Google上进行搜索运算。考虑到用户查看搜索结果的习惯和易用性，人工统计前20个搜索结果的查准率Pr@20（由于涉及到具体的搜索引擎的网页数据库，暂时不统计查全率）。实验结果如表4所示。

表4 扩展前后查准率比较示例



为了对比传统的查询扩展方法，将本文提出的方法与目前应用最广泛的伪相关反馈方法做比较。对每一个查询请求，取其前10个初始查询结果中的前5个高频实词作为扩展词加入到初始查询中，形成扩展查询请求，进而提交给搜索引擎得到最终搜索结果。同样，人工统计前20个搜索结果的准确率Pr@20，与本文提出的查询扩展方法形成对比。其中网页内容采用Java程序提取，对网页内容进行分词时采用ICTCLAS10算法，统计得到前5个扩展查询词。实验结果如表5所示。

表5 不扩展、PFR方法扩展与语义扩展的查准率比较



以Search4、5和19为例，表6列出了扩展过程和扩展结果。

表6 扩展示例

ID	初始查询	PRF扩展查询词	语义表达式	SE扩展查询
4	多边形定义	数学，初一，内角，三角形，简介	(多边形:N)['s](定义:N)	“多边形定义” OR “多边形的定义” OR “多边形概念” OR “多边形的概念”
5	MotoV3 说明书下载	手机，驱动，原装，软件，论坛	(MotoV3:N);((说明书:N)(下载:V)):VP	(“MotoV3” OR “Moto V3” OR “摩托罗拉 V3” OR “摩托罗拉 V3”) AND (“说明书下载”)
19	电子书 编写	资料，频道，化学，人员，名单	(电子书:N);(编写:V)	(“电子书” OR “e书”) AND (“编写” OR “编辑” OR “撰写”)

从表4和表5看，使用扩展查询请求提交给搜索引擎要比直接提交和PRF扩展均有一定程度的提高，这是因为扩展后的搜索请求包含了与初始查询语义相关联的更多的表达形式，而且这些表达形式不可再分，避免了分词导致的语义损失，使得准确性有所提高；而PRF扩展方法在初始查询中加入的扩展词高度依赖于初始查询结果，初始查询的不准确直接导致了无关扩展词的加入，反而影响了查询的性能。但是可以看出，使用扩展查询请求也并非万能，一方面实验用的语义单元表示库的规模不是很大，覆盖面有限；一方面用户提交的查询请求太过简单，无法从中提取用户感兴趣的领域；另一方面是由于搜索引擎从根本上讲是基于关键字的，缺乏知识理解的能力。这是当前搜索引擎普遍存在的问题，也是我们后续研究的主要方向。

总结 本文利用语义单元的语义表达能力和语义单元之间的语义关系，对搜索引擎用户的初始查询请求进行扩展。实验证明，使用扩展查询请求进行检索，得到的查准率明显好于未扩展时的查准率。本算法的时间复杂度是$O(L)$，扩展所消耗的时间只与初始查询的长度L有关，与语义单元表示库的规模无关，这对实时性要求比较高的搜索引擎是非常有意义的。

参考文献

1 Sparck J K. Automatic Keyword Classification for Information Retrieval. London: Butterworths, 1971

2 Deerwester S, Dumai S T, Furnas G W, et al. Indexing by latent semantic analysis. Journal of ACM Transactions on Information Systems, 2000, 18(1): 79～112

3 Jing Y, Croft W B. An association thesaurus for information retrieval. In: Proceedings of the Intelligent Multimedia Information Retrieval Systems, 1994. 146～160

4 Qiu Y, Freib H. Concept based query expansion. In: Korfhage R, Rasmussen E M, Willett P. eds. Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1993. 160～169

5 Rocchio Jr J J. Relevance feedback in information retrieval. In: Salton G. ed. The SMART Retrieval System: Experiments in Automatic Document Processing. Englewood Cliffs, New Jersey: Prentice-Hall, 1971. 313～323

6 顾榕，王小平，曹立明. 一种基于潜在语义分析的查询扩展算法. 计算机工程与应用，2004 (18)：23～25

7 岳文，陈治平，林亚平. 基于查询扩展和分类的信息检索算法. 系统仿真学报，2006，18(7)：1926～1929

8 崔航，文继荣，李敏强. 基于用户日志的查询扩展统计模型. 软件学报，2003，14(9)：1593～1599

9 Navigli R, Velardi P. An analysis of ontology-based query expansion strategies. In: Proceedings of the 14th European Conference on Machine Learning, Workshop on Adaptive Text Extraction and Mining, Cavtat Dubrovnik, Croatia, 2003

10 GAO Qingshi, HU Yue, LI Li, et al. Semantic language and multi-language MT approach based on SL. Journal of Computer Science& Technology, 2003, 18(6): 848～852

11 高小宇，高庆狮，胡玥，等. 基于语义单元表示树剪枝的高速多语言机器翻译. 软件学报，2005，16(11)：1909～1919