一种改进的基于关系的信息检索技术*)

李 岩[1] 文 健[1] 李舟军[2]

(国防科学技术大学计算机学院 长沙 410073)[1] (北京航空航天大学计算机科学与工程学院 北京 100083)[2]

摘 要 有研究工作表明现有的基于关系的信息检索技术(RIR)优于基于项(term)或基于语义(concept)的IR技术,但仍存在显而易见的缺陷,即不能明确关系本身,只能表达概念A,B是存在关系的概念对。本文提出一种改进的基于关系的IR技术—IRIR (Improved Relation-based Information Retrieval),就是要明确关系的取值和属性,整合概念对和关系的信息为三元组表达式(triple),通过以下匹配方法获取未知信息。对于文本中出现的知识表示为R(relation)[First Concept,Second Concept],对于疑问代词(如what)开头的查询表达为R(relation)[First Concept,Unknown],对于疑问副词(如how)开头的查询表达为R(Unknown)[First Concept,Second Concept],当文本与查询的三元组表达式中已知部分匹配一致时,则得到查询未知部分的一个取值。由此,既可以实现类似QA(query answer)功能,又可以完成精确信息检索。基于Drexel大学DM & Bioinformatics Lab开发的生物医学文献搜索引擎(2004版,简称为RIRS),我们开发了一个能实现IRIR技术和功能的实验IR引擎—IRIRS (Improved Relation-Based IR System),该系统使用UMLS和WordNet两大权威本体库分别确定概念和关系,在博士入学考试英语阅读理解测试集上的实验结果令人满意,IRIRS将文字段级别的检索精确度MA PP(Mean average passage precision)从RIRS的64.44%提高到74.28%。这表明,在IR中应用改进的基于关系的信息检索技术是非常具有探索价值的。

关键词 信息检索,关系抽取,查询分析,三元组结构

Improved Relation-based Information Retrieval Technology

LI Yan[1] WEN Jian[1] LI Zhou-jun[2]

(School of Computer,National University of Defense Technology,Changsha 410073,China)[1]

(School of Computer Science,Beijing University of Aeronautics & Astronautics,Beijing 100083,China)[2]

Abstract One of the limitations with the traditional relationship-based IR methods is that a relation is often recorded as a binary form,such as R(First Term,Second Term),which is only composed of general information of a pair of two terms which are semantically and syntactically related to each other. To tackle this problem,we explore an improved technique by using of triples in information retrieval for precision-focused biomedical literature search. In this paper,a triple is defined as a data structure for the integration of a pair of concepts as well as a verb phrase or sometimes a special noun we extract from the sentence as the relation of the above concepts pair,and stores relation and concepts information. Unlike the traditional relationship-based model,our model represents a document or a query by a set of triples,such as R(relation)[First Concept,Second Concept]. Since some semantic and syntactic exceptions occur in documents and queries,the different types of triple should be permitted,e.g. a query:"What does the mad cow disease come from?" has a triple:R(come from)[First Concept(mad cow disease),Unknown]. Therefore,we can get the "answer" of the unknown thing in query if some documents have the matching triples in the index. Of course,we will apply the advanced ontology-based approach to extract generic concepts and their relations by using both UMLS and WordNet,and we have implemented a new approach to rank retrieved passages from same or different documents corresponding to measuring system performance protocol in TREC 2007 Genomics Track. A new version (we called it IRIRS) of the relation-based IR system which has been developed by DM & Bioinformatics Lab of Drexel University in 2004 (we called it RIRS),is then built for the improved relation-based search in the area of biomedical literature IR and DM. We use IRIRS to improve the retrieval result of tests of English reading comprehension. The experiment shows promising performance of relation-based IR. Mean average passage precision (MAPP),the character-based MA Pmeasuring passage-level retrieval performance,for 64 topics is significantly raised from 64.44 % (the result of RIRS) to 74.28%. Furthermore,the experiment shows more expressiveness of relation and triple structure for the representation of information needs,especially in the area of biomedical literature.

Keywords Relationship-based information retrieval,Relation extraction,Query parsing,Triple integration

在信息技术蓬勃发展的今天,现代信息检索技术(modern information retrieval technology)在广泛的需求和多种技术平台的支持下迅猛发展。海量信息的产生使信息检索(information retrieval,简称IR)、数据挖掘(data mining,简称

*)本文受到国家自然科学基金项目(60573057,60473057 ,90604007)的资助。李 岩 硕士生,主要研究领域为基于关系的信息检索技术;李舟军 博士,教授,博士生导师,主要研究方向为数据挖掘与生物信息学、进程代数理论、安全协议的形式化验证;文 健 博士研究生,主要研究领域为信息检索和文本挖掘。

DM)技术的发展和应用成为必然。近几年来,美国等发达国家政府在生物医学领域的投入与日俱增,NCBI(National Center for Biotechnology Information)、UMLS(Unified Medical Language System)、TREC(Text Retrieval Conference)等一些生物医学文献挖掘相关的科研、测试平台也日趋完善,有力促进了生物医学文献挖掘(biomedical literature data mining technology)领域的发展。TREC 的推动使得 IR 成为有标准、有组织、有目标的领域,促使现代 IR 技术向更精确、智能化、自动化程度更高的方向发展。因此,将研究工作与 TREC 联系起来,将 IR 技术与生物医学工程联系起来,可以让成果更有说服力,也更实用。

1 简介

传统的 IR 系统在分析处理用户查询(query)时过多依赖于语料库(corpus)或本体库(ontology)[8-10],忽略了用户的查询本意,而这些往往是用户"隐式"提供的,即查询提问的未知信息,一般用"属性"和"关系"来定位。多数 IR 系统都是按照用户查询中的"显式"信息或其扩展信息进行检索和排序,如图 1。未充分提取用户"隐式"提供的信息会大大影响 IR 系统的性能,只有将语料库信息与查询本意信息相结合,检索才是高效的。

Query: Which test would be best for diagnosis of *ovarian cyst* in this case?
Query Term: *ovarian cyst* ("显式" 信息)
User's Real Need: *Unknown* thing, which is a test, has relation(*diagnosis*) with *ovarian cyst*. ("隐式" 信息)
Result:

Document rank	Terms appearing in the document
Document1	many "*ovarian cyst*"
Document2	many "*ovarian cyst*", "*culdocentesis*"(one of the diagnosis methods of ovarian cyst), "*pelvic ultrasound*"(one of the diagnosis methods of ovarian cyst)
…	…
DocumentN	few "*ovarian cyst*", many "*culdocentesis*",…, "*culdocentesis is the best test for diagnosis of ovarian cyst*"

图 1 多数 IR 系统按照用户查询中的"显式"信息进行结果搜索和排序

传统的 IR 用词汇项(term)表示文本[19],后来改进为基于语义的方法,用概念(concept)表示文本[20]。近年来的基于关系的 IR 技术[2]有了重大突破,考虑了概念间的关系,把 IR 技术提高到新的层次。但是这种被称为基于关系的 IR 技术不是真正意义上的"基于关系",因为系统根本不能提取出"关系"本身。美国 Defense Advanced Research Projects Agency(简称 DARPA)与 National Institute of Standards and Technology(简称 NIST)共同举办的文本检索会议(Text REtrieval Conference,简称 TREC)为研究生物医学领域上的信息检索技术,自 2003 年开始举办 Genomics Track,为各类检索系统提供测试机制,以比较它们在生物医学领域的检索效能,已成为近年来热门的测试项目。TREC 2006 Genomics Track 给出的官方文件[1]中,明确提出 IR 领域的关系可以是"任何值,一般为动词",并且将该届测试的所有问题表示为图 2 式(1)。将该格式去特殊化后推广到一般的文本表示,见图 2 式(2,3),可以表示所有查询问句和文本,并通过以下匹配原理获取查询提问的未知信息,见图 3。当文本与查询的三元组表达式中已知部分匹配一致时,则得到查询未知部分的一个取值。由此,既可以完成高精度信息检索,又可以实现类似 QA(query answer)功能[1)]。

Biological object (1..many) <--relationship--> Biological process (1..many) ...(1)
First Concept(one) <--relationship(atomic)-->Second Concept(one) ...(2)
R(relation, passage())[First Concept(CUI,TUI,STR), Second Concept(CUI, TUI,STR)] ...(3)

图 2 文本和查询问句的基于关系的统一三元组表示模式

文本: R(*relation*, passage(passage1,passage2,...,passageN))
[*first concept A*, *second concept B*]
查询 1: R(*relation*, passage(?, ?, ...))
[*first concept A*, Unknown[1]]
查询 2: R(Unknown[2],passage(?, ?, ...))
[*first concept A*, *second concept B*]
结果: passage(passage1,passage2,...,passageN); Unknown[1] = *second concept B*; Unknown[2] = *relation*.
注释: 查询 1: 疑问副词(如 how)开头的查询; 查询 2: 疑问代词(如 what)开头的查询。

图 3 获取查询中的未知信息的匹配原理

本文第 2 部分论述基于关系的查询和文本表示,第 3 部分论述概念和关系的抽取,第 4 部分介绍对应于不同的查询、检索的相关性判定和排序方案。第 5 部分是对实验的设计和结果分析,最后是对研究工作的总结。

2 基于关系的查询与文本的表示

2.1 查询与文本

查询具有多样性:1)项查询(term query),仍在普遍使用,由多个(>=1)查询项(term)组成。2)问句查询(question query),近年来应用的趋势,很多解释叙述型查询都可以转化为问句形式[1],其特点是以疑问代词 what 或疑问副词 how 为首特殊疑问句为主[2)],其他特殊疑问句可以归并到这两种情况[12]。3)陈述句查询(declaration query),与文本语句相似。文本具有单一性,基本由陈述句构成,因此可以按 3)处理,见表 1。

2.2 基于关系的查询与文本语句的表示

2.2.1 三元组的表示模式

T(CUI=?, TUI=?, STR=?, Attribute=?, Location=?, passage())
R(RCUI=?, @RCUI=?, STR=?, Location=?)
R(RCUI=?, @RCUI=?, STR=?, value=?, passage())[T1(CUI=?,...), T2(CUI=?,...)]

图 4 概念、关系、三元组的表示格式

图 4 中 CUI (Unique Identifier of Concept)、TUI(Unique Identifier of Semantic Type)是概念及其语义类型的唯一识别码[3)],STR(String)字符串记录,Attribute 概念属性的记录,

1) http://trec.nist.gov/data/qa/2006_qadata/aq.06.guidlines.html
2) http://ir.ohsu.edu/genomics/2006data.html topics
3) http://www.nlm.nih.gov/research/umls/meta2.html

Location 是概念或关系在文本中的偏移量，用以定位概念对(concept pair)之间的关系，见表 3；关系是原子的(atomic relation)，即其中无并列成分[7]，是保证完整语义的最小单位，RCUI(Unique Identifier of Relation)、@RCUI(Unique Identifier of Hypernymy of Relation)是 WordNet 给出的关系及其上位词的唯一识别码，对于查询中不涉及无关系的情况(见表 1Query(a,b,e))，RCUI 赋值为"00000001"，value 是三元组在语料库中出现的次数。

2.2.2 查询与文本的表示模式

表 1 查询与文本的类型及表示(其中需要检索的未知信息记为 Unknown，表达式中未列出的参数取值为空)

Type	Category	Example
Term Query	Word Query	Query(a)：aids patient cancer capoci sarcoma. R(RCUI="00000001",value)[ConceptList(CUI,TUI,STR)]
	Concept Query	Query(b)：C1418941 (PRNP) C0085209(Mad Cow Disease) R(RCUI="00000001",value)[ConceptList(CUI,TUI,STR)]
Question Query	"How" Query	Query(c)：How do Cathepsin D (CTSD) and apolipoprotein E (ApoE) interactions contribute to Alzheimer's disease? (RCUI=Unknown,@RCUI,STR="How",value) 9[T1(CUI,TUI,STR),T2(CUI,TUI,STR)]
	"What" Query	Query(d)：What is the role of IDE in Alzheimer's disease? R(RCUI,@RCUI,STR,value) [T1(CUI,TUI,STR),T2(CUI= Unknown,TUI,STR= "What")]
Declaration Query or Text	Description Sentence	Query(e)：Find all reports describing mouse peptidoglycan recognition proteins (PGRP). R(RCUI="00000001",value)[ConceptList(CUI,TUI,STR)]
	Statement Sentence	Query_Text(f)：HIV patients suffering from the Capoci sarcoma tumor. R(RCUI,@RCUI,STR,value)[T1(CUI,TUI,STR),T2(CUI,TUI,STR)]

3 概念与关系的抽取与整合

3.1 概念抽取

生物医学文献和相关查询中出现的名词和名词短语几乎都是专业词汇。UMLS 的超级叙词表(Metathesaurus)中收录了 200 万个生物医学及相关领域的概念，几乎覆盖了这些文献中涉及的所有概念或名词词汇[3]。在生物医学文献检索中使用 UMLS 抽取的概念和名词项[11,17,18]不会出现歧义[15]，工作流程如图 5。对于代词的指代消解和词义消歧分别应用美国 Drexel 大学数据挖掘与生物信息实验室(DM & Bioinformatics Lab)开发的启发式方法[2]和改进的 Lesk 方法[3]。缩写词处理采用 Dimitrov 小组开发的方法[4]。

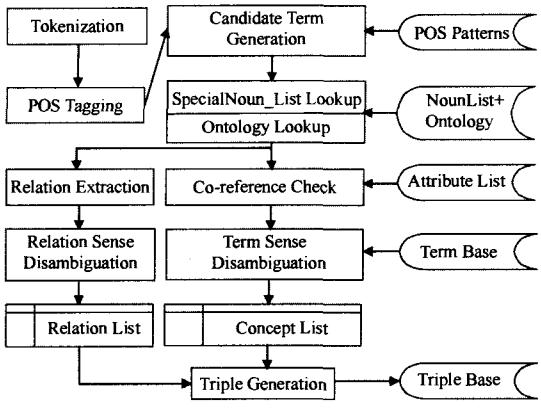

图 5 概念和关系的抽取和三元组生成流程

3.2 关系抽取[15]

这里的关系抽取极富挑战性。IR 的关系抽取有别于 IE(Information Extraction)领域[21]。TREC 提出 IR 领域的关系可以是"任何值，一般为动词"[1]，这显然太模糊。而文献[2]将概念抽取完成后直接生成关联概念对并视其为关系，这种只明确关系涉及的概念，而不明确概念间有何种关联的思路虽易于实现，却有很大的缺陷。在英语词典[5]中，关系是定义在至少两个事物上的，用以表达某种关联性的词语。由此可以认为动词直接体现了概念间的关系，另外我们也将考虑某些具有动词涵义的特殊名词。

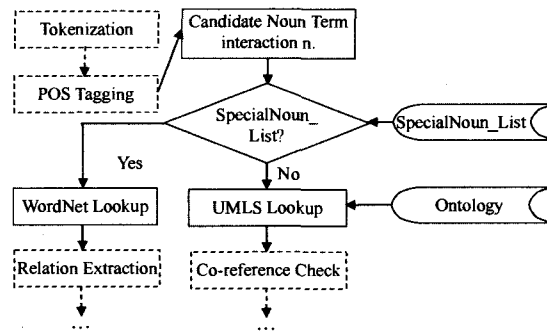

图 6 名词项处理子流程

3.2.1 抽取关系

动词和一些特殊名词是关系的载体。文本通过分词、词性标注后就能得到动词和名词。名词在 UMLS 中查找(Ontology Lookup)前，先与特殊名词列表(Special Noun List)中的名词比对，若是该列表中的名词，直接转换为该名词对应的动词(SpecialNoun_List 实现转换功能)进行下面的处理，如 interaction n. 转换为 interact v，见图 6。关于特殊名词表，我们只是提出了初步设想，建立了机制，并没有形成一个完整的词汇表。目前表中都是在处理查询时遇到的或某些常见的具有"动词感"的名词。

3.2.2 表示关系

WordNet 与 UMLS 不同，它收录了足够多的普通英语词汇。更值得一提的是 WordNet 将所有词汇按同性同义词分

类，汇集成 synset(synonymy set)，根据 synset 在数据库文件中的位置(location in database file)编号。同一词性的 synset 编号唯一，动词 synset 共 13，650 个，编号从 00001740～02746781，非连续，记为 RCUI。WordNet 为关系提供了类似于概念的语义类型的上位关系(hypernymy：is a kind of…，记为@)，即 A 是 B 的上位关系⇔B is a kind of A。上位关系为我们处理关系未知的查询问句提供了有效的方法，如例 1。虽然关系未知但关系的上位关系已知，显然有利于未知关系的定位。这样既方便检索，又能有效提高精度。

Example 1　Topic：How does nucleoside diphosphate kinase (NM23) contribute to tumor progression?

Query：R (Unknown, @ = contribute) [A (nucleoside diphosphate kinase (NM23)), B(Tumor progression)]

关系和概念的 Location 主要用于组合关系和概念对，Location 的值为关系和概念出现在语料库文本中的偏移量，见表 3。表 3 说明存在一个关系，位于语料库中编号为 15485830 文件的第 6 段 (在文本中起始于第 4305 个字符)中的第 5 个句子(在段落中起始于第 1102 个字符)，并起始于该句的第 20 个字符(包括空格)。

表 2　关系的偏移量

文件编号	段落偏移量[1]	句子偏移量	3 位句中偏移量	关系的偏移量 (Location)
15485830	4305	1102	20	(4305+1102) * 1000+20=5407020

3.2.3　关系“消歧”

据 WordNet 统计，动词词条中有 45.5%的多义词(共有 11488 个动词，5227 个多义词)[4]。当得到的动词是多义词时，如何取舍多个词义，成为亟待解决的问题。对此我们采取一种简易方法：将多义动词的第 1 个词义作为关系的概率词义(该词义出现在指定语料库[5]中的次数最多)。由此可以获得较高的词义覆盖率，并大大降低计算的复杂度。当然，只用 1 个词义标识关系毕竟不太精确，3.3 节中将做进一步讨论并提出解决方案。

3.3　概念与关系整合为三元组

概念与关系整合为三元组的工作原理和算法见算法 1。首先计算概念对的中心位置(见算法 1 中式(1))。关于概念配对，我们采用文献[2]关于实体-实体关系生成的方法。概念对(Concept Pair)与关系列表(Relation List)中第 n 个的关系组合形成三元组，即表明该关系与概念对在同一个句子中(见算法 1 的 Step3)，并且距离概念对的中心位置最近(见算法 1 中式(2))，n 值的计算见算法 1 中式(3)。在得到 triple 三个分量的值以后，就可以应用不同的策略界定与 triple 对应的文字段(passage)的偏移量和长度了(见算法 1 的 Step7)。文字段的界定策略有多种，本文使用的是以句子为最小单位组成文字段的策略，即文字段的起始(offset)是关系所在的句子的偏移量，文字段的长度(length)是该句的长度。

算法 1　关系与概念对整合算法

Step1　i=0,
Step2　j=0,S is an empty set.
Step3　ConceptPair[j]. Location=
$$\frac{\text{ConceptPair[j].Concept1.Location}+\text{ConceptPair[j].Concept2.Location}}{2}\cdots(1)$$
int rl= Relation_List[i]. Location/1000, cl= ConceptPair[j]. Location/1000;
if rl==cl is true goto Step4, else Step5.
Step4　distence(j)=Abs(relation_List[i]. Location − ConceptPair[j]. Location)…(2)
S←i.
Step5　j=j+1. if j>MAX_SIZE_C goto Step6, else goto Step3. // MAX_SIZE_C is size of concept pairs list
Step6　if S is empty return false; else n=arg $\min_{j\in S}$distence(j)…(3)
Step7　offset=rl; length=length of the sentence the relation located;
passage←(document name, offset, length)
Step8　Triple Base←R(relation_List[i], passage)[ConceptPair[j]. Concept1, ConceptPair[j]. Concept2]
Step9　i=i+1. if i>MAX_SIZE_R return Triple Base, else goto Step2. //MAX_SIZE_R is the size of relation list.

WordNet 的词义非常细化，只用 1 个词义标识关系，可能会出现词义接近但不绝对相同的关系，因标识不同而无法被检索出来，以致于大大降低召回率。因此，处理查询时使用多义动词的前 2 个词义(有时也可以是两个动词的首词义)作为关系的概率词义，可将词义覆盖率提高到 90%以上。

4　相关判定—排序机制

由于最终给出的结果是被系统判定为相关的合法文字段，排序工作较以前有很大不同。相关判定-排序原则：1)匹配最佳优先原则：文本 triple 匹配上查询 triple 的分量越多排序在前；2)概念匹配优先原则：概念包含的专业信息更丰富，因此匹配上概念较多的文本 triple 排序在前，只有关系分量匹配上的不判定为相关；3)当 topic 处理为多个 triple，每个 triple 按相同权重(=1)分别进行检索。

算法 2　文本三元组的排序算法

Step1　i=0; Set 0 for rankValue of all triples in DocumentTripleList. // i is topic ID
Step2　j=0; // j is the ID of query triples of one topic
Step3　k=0; // k is the document triple ID
Step4　value= DocumentTripleList[k]. value; value_f =0, value_s =0, value_r=0; weight=1;
Step5　if QueryTripleList[j]. concept1==DocumentTripleList[k]. concept1 value_f=value * 0.4;
if QueryTripleList[j]. concept2==DocumentTripleList[k]. concept2 value_s=value * 0.4;
//查询 triple 中关系是单义关系
if QueryTripleList[j]. relation is a monosemous verb
if QueryTripleList[j]. relation ==DocumentTripleList[k]. relation && (value_f! =0 || value_s! =0)
value_r=value * 0.2;
//查询 triple 中关系是多义关系
else
if QueryTripleList[j]. relation. first==DocumentTripleList[k]. relation && (value_f! =0 || value_s! =0)
value_r=value * 0.2;
if QueryTripleList[j]. relation. second==DocumentTripleList[k]. relation && (value_f! =0 || value_s! =0)
value_r=value * 0.2; weight=0.4
DocumentTripleList[k]. rankValue = DocumentTripleList[k]. rankValue+(value_f+value_s+value_r) * weight;…(1)
Step6　k++, if k>MAX_SIZE_DT goto Step7, else goto Step4. // MAX_SIZE_DT is the size of document triples list.
Step7　j++, if j>MAX_SIZE_QT goto Step8, else goto Step3. // MAX_SIZE_DT is the size of query triples list.
Step8　Integrate the different triples in the same passage (offset, length), and accumulate the values.
Step9　passages are submitted for Topic[i]: ranking the passage from the DocumentTripleList, starting with top-ranked passage which has highest value, and preceding down to the passage of lower value as many as 1000.
Step10　i++. if i>MAX_SIZE_TL return true, else goto Step2. // MAX_SIZE_TL is the size of topics list.

相关判定-排序(judging_ranking)算法见算法 2。将第 k 个文本 triple 比对第 i 个 topic 的第 j 个查询 triple，根据匹配

[4] http://wordnet.princeton.edu/man/wnstats.7WN

[5] Brown 语料库(当代美国英语标准语料库)的语料和一个中短篇小说的全文

上的分量类型和数量打分，概念分量匹配上多的分值较高(=文本 triple[k]分值 * 加权值 0.4，文本 triple 分值即为该 triple 在语料库中出现的次数)，关系分量匹配上的分值较低(=文本 triple[k]分值 * 加权值 0.2)；对于查询 triple 中关系是多义的情况，首词义(a)或次词义匹配上时文本 triple 的排序计算公式(见算法 2 中式(1))的加权值(weight)分别为 1 和 0.4，对于关系是单义的情况，排序计算同(a)；只有关系分量匹配上的分值为 0；合并处在同一文字段的不同 triple，累加其排序权值(rankValue)，最后按分值高低提交 triple 对应的文字段(即 triple 表达式中的 passage 部分)。

5 实验

5.1 实验使用的搜索引擎和测试集

基于 Drexel 大学 DM & Bioinformatics Lab 开发的生物医学文献搜索引擎(2004 版，简称为 RIRS)，并针对 TREC 2006 Genomics Track 的需求，改进和开发了一个用于实验的 IR 引擎—IRIRS，实现了真正的基于关系的 IR 功能。新系统主要增加了文本与查询的关系抽取与消歧以及三元组格式的生成、索引和排序功能，并提交 passage 作为检索结果。测试集取自国内英语权威专家编写的博士研究生入学考试英语阅读理解辅导材料[16]中收录的自然科学类中关于生物医学的 15 篇阅读理解试题，其中不考虑结构类型(对 IR 研究没有意义)的 8 个问题(topic)，则共有 64 个问题。每个问题都附有专家编写的答案解析，便于对实验结果做出科学的分析。鉴于这套测试集的权威性，以及符合 IR 性能评估要求的问题和文本，我们认为在该测试集上进行的实验，其实验结果是可靠的。

5.2 实验的设计

改进的基于关系的系统 IRIRS 是否是一个高精确率的 IR 系统，取决于它在性能上是否优于原有系统 RIRS。在测试集上对 RIRS 和 IRIRS 系统进行对比实验，考察两个系统在文字段级别(passage level)的性能。我们将每个提问(问题主干+一个选项)作为测试的最小分划，计算 CPP(cumulative character-based precision))并最终计算 topic 的 TMAPP(Topic Mean Average Precision，参考了 TREC 2007 Genomics Track 中规定的 character-based MAP，简称 MAPP[6]，MAPP 是标识 IR 系统在文字段级别上的精确率和召回率的综合指标)。

5.3 实验结果分析

1) 实验：对 64 个 topic 依次计算 TMAPP，计算公式见图 7，RIRS 与 IRIRS 的 TMAPP 对比见图 8。式(3)计算系统提交的第 k 个文字段的累积基于字符的精确率(cumulative character-based precision，简称 CPP)，记为 $CCP\_of\_relevent\_passage(k)$，即前 k 个文字段中被标准答案标记为相关字符的字符总数(记为 $number\_of\_endcharacters\_in\_answer$)占前 k 个文字段的字符总数($total\_number\_of\_characters\_to\_submit$)，在标准答案中而系统未能检索出的文字段 CCP 是 0。(2)式计算提问 j 的所有提交的和 CCP=0 的文字段的平均精确度(average passage precision，简称 APP，记为 $APP\_of\_question(j)$)。

$$TMAPP(i)=\frac{\sum_{j=1}^{number\_of\_question} APP\_of\_question(j)}{number\_of\_questions}$$，其中 $number\_of\_questions=4; i=1,\cdots,64$ (1)

$$APP\_of\_question(j)=\frac{\sum_{k=1}^{number\_of\_relevant passage(j)} CCP\_of\_relevent passage(k)}{number\_of\_relevant passage(j)} \quad (2)$$

$$CCP\_of\_relevent passage(k)=\frac{number\_of\_endcharacters\_in\_answer}{total\_number\_of\_characters\_to\_submit} \quad (3)$$

图 7 计算 TMAPP 的公式组

2) 实验结果分析：正如我们所料，IRIRS 在文字段级别的性能优于 RIRS。主要因为 IRIRS 在 RIRS 上增加了动词(即关系)的抽取。当提问中仅出现一个概念时，RIRS 就会“盲目”地检索到大量出现这个概念但与主题不相关文字段。通常，这样的提问中会出现重要的关系信息，IRIRS 就可以利用这个 RIRS 不能识别的信息实现精确的信息检索。从图 8 和表 2 来看，IRIRS 的优势不如预期的明显，分析原因主要有：1) IRIRS 和 RIRS 都不能有效地利用形容词提供的信息，而我们使用的测试集中大量出现了涉及形容词的问题(topic)。2) 文字段级别的性能评估机制[1]不能有效区分如下类似情况：标准答案中有 1 个文字段包含位置连续的 3 个句子，系统 A 提交的相关文字段中只包含了第 1 句，系统 B 提交的相关文字段中包含了第 1、2 句，显然系统 B 优于系统 A，但这种情况下的两个系统却有相同的 TMAPP 值[7]。3) 相关文字段的界定问题有待进一步解决。比如是以句子为最小单位还是以子句或字符串为最小单位来确定相关文字段起始(offset)和长度(length)的策略各有优缺点。以上 3 点将在今后工作中解决。

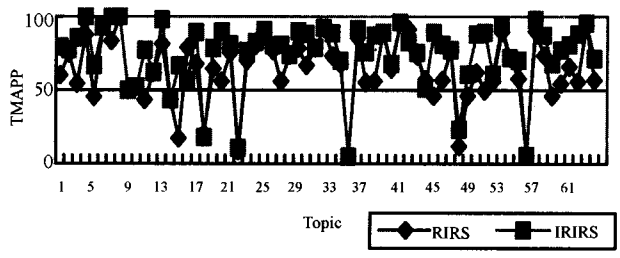

图 8 RIRS 和 IRIRS 对应于 64 个 topic 的 TMAPP(Topic Mean Average Precision)比较

结束语 本文提出了一种新的基于关系的信息检索技术，它不同于以往的 RIR 技术。我们对文本分析以建立包含关系详细信息的三元组，对查询分析以建立包含未知分量并明确未知分量属性的三元组，系统以三元组为索引，搜索相关文字段。由于三元组的格式中包含了已知或未知概念对与关系在上下文和本体库中的相关信息，很大程度上提高了自然语言处理和信息检索的智能化。未来的信息检索技术会随着计算机技术和本体技术的发展智能化程度不断提高，而基于关系的 IR 技术在 TREC 的推动下也有待完善。

[6] http://ir.ohsu.edu/genomics/2007protocol.html

[7] 详情见 TREC 2007 Genomics Track Draft Protocol 和 TREC 2006 Genomics Track Protocol 的 Measuring System Performance

表3 RIR与IRIR系统性能比较

Topic	Topic Type	RIRS_ TMAPP	IRIRS_ TMAPP	Topic	Topic Type	RIRS_ TMAPP	IRIRS_ TMAPP
[1]	推论	60	79.2	[33]	细节	73.2	88.4
[2]	细节	74.6	77.2	[34]	细节	67.6	69.8
[3]	细节	54.2	85.5	[35]	词义	5.3	4.6
[4]	推论	87.5	100	[36]	细节	85.5	91
[5]	细节	45.5	66.7	[37]	细节	55.2	75.4
[6]	细节	92.6	92.6	[38]	态度	56.2	86.9
[7]	归纳	83.3	100	[39]	细节	88.7	88.7
[8]	态度	100	100	[40]	归纳	64.2	67.7
[9]	主旨	50	50	[41]	细节	95.8	95.8
[10]	细节	52.1	52.1	[42]	细节	91.1	82.4
[11]	主旨	43.3	77	[43]	主旨	73.2	75.3
[12]	细节	61.9	61.9	[44]	词义	56.9	51.2
[13]	细节	81.3	98	[45]	细节	46	88.9
[14]	主旨	43.6	43.6	[46]	推论	56.7	80.8
[15]	细节	17.2	66.6	[47]	细节	77	77
[16]	词义	78.9	56.3	[48]	细节	11.7	23.2
[17]	细节	67.5	89	[49]	主旨	45.9	60.5
[18]	细节	20.4	17.8	[50]	主旨	62	87.7
[19]	主旨	64.6	78.4	[51]	细节	49.6	88.9
[20]	细节	56	89.9	[52]	细节	55.9	60.4
[21]	推论	75.1	80.7	[53]	词义	89.2	94
[22]	词义	8.7	10.4	[54]	细节	71.7	71.7
[23]	细节	69.6	76.2	[55]	细节	57.8	69.9
[24]	细节	80.1	82.5	[56]	词义	5.5	5.5
[25]	细节	89	90.7	[57]	归纳	89.8	97.6
[26]	细节	76.7	80.1	[58]	词义	73.4	86.9
[27]	细节	56	80.4	[59]	细节	46	67.8
[28]	细节	73.3	74	[60]	主旨	54.3	76.9
[29]	细节	78.3	89.7	[61]	细节	65.9	80.4
[30]	细节	66.6	87.9	[62]	细节	55.5	87.6
[31]	细节	78.6	78.6	[63]	推论	94	95.2
[32]	细节	92.1	92.1	[64]	细节	54.7	71.3
		RIRS				IRIRS	
MAPP		64.44%				74.28%	

致谢 在此，向曾经对本文提出宝贵建议的专家以及曾参与本文内容讨论的所有老师、同学表示衷心的感谢。

参考文献

[1] Hersh W, Cohen A M, Roberts P, et al. TREC 2006 Genomics Track Overview// Proceedings of the 15th Annual Text Retrieval Conference National Institute of Standardsand Technology. 2003

[2] Zhou Xiaohua, Hu Xiaohua, Lin Xia, et al. Relation-based document retrieval for biomedical literature databases. DASFAA - Database Systems for Advanced Applications. Singapore, 2006

[3] Lesk M. Automatic Sense Disambiguation: How to Tell a Pine Cone from and Ice Cream Cone// Proceedings of the SIGDOC'86 Conference. ACM, 1986

[4] Dimitrov M, Bontcheva K, Cunningham H, et al. A Light-weight Approach to Coreference Resolution for Named Entities in Text // Proceedings of the Fourth Dis-course Anaphora and Anaphor Resolution Colloquium (DAARC). Lisbon, 2002

[5] Hordby A S, Li Beida. Oxford Advanced Learner's English-Chinese Dictionary. Forth Edition. The Commercial Press, Oxford University Press

[6] WordNet 2.1 Copyright 2005 by Princeton University. WordNet \2.1\dict\data.verb

[7] Ding J, Berleant D, Xu J, et al. Extracting Biochemical Interactions from MEDLINE Using a Link Grammar Parsern // 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'03). 2003

[8] Cohen W W, Singer Y. Simple, Fast, and Effective Rule Learner // Proceedings of the Sixteenth National Conference on Artificial Intelligence and Eleventh Conference on Innovative Applications of Artificial Intelligence. July 1999: 335-342

[9] Mitra C U, Singhal A, Buckely C. Improving Automatic Query Expansion// Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1998: 206-214

[10] Robertson S E. On term selection for query expansion. Journal of Documentation, 1990, 46: 359-364

[11] Mooney R J, Bunescu R. Mining Knowledge from Text Using Information Extrac-tion. SIGKDD Explorations (special issue on Text Mining and Natural Language Processing), 2005, 7(1): 3-10

[12] 薄冰，何政安. 薄冰英语语法. 2005新版. 北京：开明出版社，1998: 390-392, 480-492

[13] Ponte J M, Croft W B. A Language Modeling Approach to Information Retrieval// Proceedings of the 21st annual international ACM SIGIR conference on Research and De-velopment in Information Retrieval

[14] Salton G, Wu H, Yu C T. The measurement of term importance in automatic indexing. Journal of the American Society for Information Science, 1981, 32(3): 175-186

[15] Sanderson M. Word sense disambiguation and information re-trieval// Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1994: 142-151

[16] 吴永麟，习天辉. 博士硕士生入学考试英语阅读精粹. 2006版. 北京：学苑出版社，2006: 72-118

[17] Soderland S, Fisher D, Aseltine J, et al. CRYSTAL: Inducing a Concep-tual Dictionary// Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. 1995: 1314-1319

[18] Soderland S. Learning Information Extraction Rules for Semi-structured and Free Text. Machine Learning. 1998, 34: 233-272

[19] Mitra C U, Singhal A, Buckely C. Improving Automatic Query Expansion// Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1998: 206-214

[20] Leroy G, Chen H. Meeting Medical Terminology Need-The Ontology-Enhanced Medical Concept Mapper. IEEE Transactions on Information Technology in Biomedicine, 2001, 5(4): 261-270

[21] Palakal M, Stephens M, Mukhopadhyay S, et al. A multi-level text mining method to extract biological relationships // Proceedings of the IEEE Computer Society Bioinformatics Conference (CBS2002). Aug. 2002: 97-108