本体匹配中基于词义组合的词法分析算法

刘秀磊[1,2],廖建新[1,2],朱晓民[1,2],杨 迪[1,2],徐 童[1,2]
(1.北京邮电大学网络与交换技术国家重点实验室,北京 100876;2.东信北邮信息技术有限公司,北京 100191)

摘 要: 针对本体匹配中相似性词法分析算法的不足,提出一种基于词义组合的词法分析算法.该算法首先通过 WordNet 发现本体中单词的合适词义,并扩展它们,然后基于本体里的语义元素形式化的定义实体的词法信息标记,最后推理出实体词法信息间的包含关系.针对一组工业本体的测试结果表明该算法有助提高系统的覆盖率.

关键词: 本体匹配; 词法分析; WordNet; 包含关系
中图分类号: TP182 **文献标识码:** A **文章编号:** 0372-2112 (2012) 08-1624-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2012.08.020

Lexical Analysis Based on Combining Senses in Ontology Matching

LIU Xiu-lei[1,2], LIAO Jian-xin[1,2], ZHU Xiao-min[1,2], YAND Di[1,2], XU Tong[1,2]
(1. *State Key Lab of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing* 100876, *China*;
2. *EBUPT Information Technology Ltd., Beijing* 100191, *China*)

Abstract: The paper presents a lexical analysis algorithm based on combining senses of words for computing subsumption relations between entities in ontology matching. It firstly finds the suitable sense of each word and extends it; then formally defines the representation of an entity notion based on semantic elements; finally, infers subsumption relations between all possible pair of entities, one from each of two ontologies. The experiments, over four real in use ontologies, show that the algorithm helps to increase the recall of the system.

Key words: ontology matching; lexical analysis; WordNet; subsumption

1 引言

随着本体使用的日益增长,表示相似(或相同)领域共享概念模型的大部分本体往往是由不同背景知识的工程师使用各种术语构造和维护.这些表示相似(或相同)领域的不同本体之间的异构性阻碍了系统对知识的共享、重用和互操作.本体匹配则是解决本体异构问题的方法之一.

目前大部分本体匹配系统[1~3]在分析本体的词法信息时多采用基于文本的相似性方法(比如文献[5~6])或基于词典的相似性方法(比如文献[7~8])计算不同本体中实体之间的词法相似性.通常这些相似性的值是在[0,1]之间的实数[4],不包含任何的语义关系,并且大部分本体匹配系统在采用相似性算法时忽视了以下问题:

• 词义组合性问题:实体标签和评论由多词汇构成,因此需考虑词汇间的关系.假设有两个概念标签 *Book* 和 *BookTitle*,尽管它们之间存在一定的相似性,但实际上 *Book* 与 *BookTitle* 表示不同的概念.再比如概念评论"*monograph or collection*"的词法信息表示的是"*monograph* ∪ *collection*".

• 词义模糊性问题:在词典里没有关系的词义,可以表达相同的概念.假如有两个属性 < *Book*, *publishedBy*, *IEEE* > 和 < *Book*, *hasPublisher*, *IEEE* >,尽管来自"*published*"和"*publisher*"的任何词义都不相关,但它们表示相同的角色,即出版商.

• 词义难寻问题:根据本体的上下文获得单词的合适词义在自然语言处理领域依然是个挑战.因此目前的词义发现技术会导致部分单词得到错误的词义.显然,得到错误的词义之后再计算它们之间的词义相似性是不可靠的.

本文提出的基于词义组合的词法分析算法,它不完全依靠单词的合适词义,通过使用 *WordNet* 里特定的词法关系,扩展单词的合适词义到一组词义以探索词义模

收稿日期:2012-01-28;修回日期:2012-07-10
基金项目:国家自然科学基金(No.61121001,No.61072057,No.60902051,No.61101119);长江学者和创新团队发展计划基金(No.IRT1049);国家科技重大专项基金(No.2011ZX02002-001-01,移动互联网总体架构研究)

糊性问题和词义难寻问题.此外,它通过组合单词的多个词义以及本体里的语义元素(例如∪和∩)形式化的定义了实体的词法信息标记以探索词义组合性问题.该算法有助于发现实体间潜在的匹配.

2　相关工作

随着语义 Web 和大量基于本体应用的发展,本体匹配已经成为目前的研究热点之一.文献[1~3]调查了各种本体匹配系统,并从不同方面对它们进行归类,也提出在本体匹配中可利用的信息,包括:词法信息、结构信息、语义信息、外部数据信息和个体信息.通常不同的本体匹配系统通过各种信息技术使用上述中的一种或多种信息完成匹配.

这些技术包括系数计算(coefficient computation)[5],图匹配(graph matching)[10],合成理论(hybrid methods)[7],机器学习(machine learning)[11,12],马尔科夫网(Markov network)[6],向量空间模型(vector space models)[13],优化技术(optimization techniques)[14],贝叶斯决策论(Bayesian decision theory)[15],以及推理机制(reasoning mechanisms)[5,8]等.

本文仅关注本体匹配系统利用词法信息的技术,关于匹配过程中的语义信息请参考文献[16].目前本体匹配系统多采用基于文本相似性的方法和基于词典相似性的方法计算本体中实体间相似性以表达词法信息.

基于文本相似性的方法[5,6,8,10,13~15]又可分为基于字符串的方法和基于描述文档的方法.基于字符串的方法是较常使用的一类技术,主要有计算编辑距离的方法、计算单词前后缀相似性的方法、计算 Jaro-Winler 分数的方法等.基于描述文档的方法将实体的相关文本视为文档,通过对不同实体的文档进行相似性计算,从而判断实体间是否存在词法关系.通常,基于文档的方法不仅仅与实体的词法信息有关,也会涉及一些实体的结构信息(或分层信息).

基于词典相似性的方法[7,8,13~15]首先使用自然语言处理技术(比如词干提取、去除停用词、分词等)将实体的名称以及评论等当作自然语言文本来处理,然后借用外部语言资源来发掘实体间的相似性以提高匹配性能.可借用的语言资源包括词典、领域叙词表和术语表等.目前使用较多的是 WordNet.它们通过利用词表对同级关系和层级关系的描述(比如同义关系、反义关系等)来提高对本体中同形不同义或同义不同形的异构现象的处理能力.

为了获得更好的匹配效果,很多匹配系统也综合基于文本相似性与基于词典相似性方法计算不同实体间词法信息的相似性[8,13~15].它们首先计算文本相似性(基于字符串的方法或基于描述文档的方法),然后基于外部词典 WordNet 等计算词义相似性,最后通过加权累加多种相似性以完成对词法信息的分析.它们之间的不同在于使用了不同的方法计算文本相似性和词义相似性,以及采用不同的方式结合文本相似性和词义相似性.例如,ILIADS[8]首先使用 Jaro-Winkler 分数法计算实体间的文本相似性,然后基于 WordNet 计算词义相似性,最后使用文本相似性和词义相似性的最大值作为最终的词法相似性.

本文提出基于词义组合的词法分析算法.该算法并未通过计算实体间的相似性分析词法信息,它扩展了单词的单个合适词义到一组词义,探索词义难寻问题和词义模糊性问题.它也通过构造实体词法信息标记的方式探索词义组合性问题,并基于实体标记(请参阅 4.3 节)推理出实体间的词法关系(比如⊆、⊇和≡).

3　系统架构

本节简述了使用基于词义组合的词法分析算法的 OACLEI 系统的架构(如图 1 所示).该系统主要分为两个阶段:词法分析阶段(图 1 竖虚线左侧所示)和语义分析阶段(图 1 竖虚线右侧所示).本文主要涉及词法分析阶段.OACLEI 输入两个本体 O^1 和 O^2,输出一组来自不同本体的实体间的匹配.本文使用两个样例本体解释各种示例,它们分别来自于 Ontology Alignment Evaluation Initiative (OAEI) 2009 标准测试集的 101 文件夹和 302 文件夹.为了表示样例本体中的实体,采用 < 101 (或 302):实体标签 > 的方式.

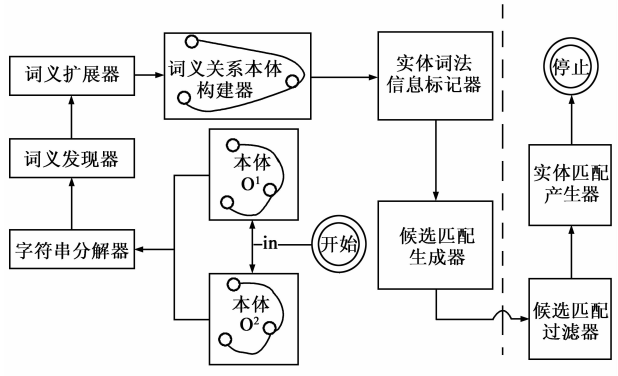

图1　OACLEI本体匹配系统架构图

词法分析阶段主要包括 6 个组件(如图 1 竖虚线左侧所示).该阶段首先将本体里实体标签和评论中的字符串分解成合理单词(字符串分解器,请参阅文献[9]),例如实体标签是 *CollectionList*,分解结果则是 < *Collection*, *List* > ;然后为这些单词找到合适的上下文词义(词义发现器,请参阅文献[4]);基于 *WordNet* 里特殊的词法关系,扩展单词的合适词义到一组词义(词义扩

展器,请参阅 4.1 节),该组词义表示了本体里的单词在目前的上下文中可能的意思,从而克服了仅发现单词单个词义的缺点,这有助于发现实体间潜在的匹配.词法分析阶段在获得输入本体中单词合适词义及其扩展后,需要寻找任意两个来自不同本体的词义之间的语义关系,并基于此构建词义关系本体,以便在推理候选匹配时使用(词义关系本体构建,请参阅 4.2 节);最后通过组合单词的多个词义以及本体里的语义元素(例如∪,∩)形式化的定义实体词法信息的标记以表达该实体在本体里的词法信息(实体词法信息标记器,请参阅 4.3 节),并推理出候选匹配(候选匹配生成器,请参阅 4.4 节).候选匹配集合是语义分析阶段的基础.

语义分析阶段(如图 1 竖虚线右侧所示)的主要目的是分析本体里的构造器和描述逻辑公理,探索本体里所蕴含的语义信息.该阶段首先过滤前一阶段产生的候选匹配集合,然后基于语义信息生成来自不同本体的实体间的匹配.

4 基于词义组合的词法分析算法

本节分析了本体中实体标签和评论所蕴含的词法信息.它首先扩展单词的合适词义,然后形式化的定义了实体标记用以表示本体里实体的词法信息,最后基于实体标记推理出本体匹配中实体间的候选匹配,为语义分析阶段提供基础.

4.1 词义扩展

为单词定义合适的词义(suitable sense)并不意味着在匹配时仅使用单词的单个词义,仍需进一步挖掘 WordNet 所提供的信息,以便找到该单词在本体的上下文中可能的词义.因为仅根据 WordNet 和本体提供的信息决定哪个词义最适合本体里的单词依然是个挑战.例如在节 1 中的例子,"*published*"和"*publisher*"之间的任何词义都不相同,但它们都表示相同的角色.图 1 中的词义扩展器用于解决这类问题.

表 1 词法关系和类公理之间的转换规则

WordNet Relation	Symbol	ClassAxiom
Hypernym	⊇	SupClass Axiom
Hyponym	⊆	SubClass Axiom
Holonym	⊇	SupClass Axiom
Antonym	⊥	DisjointClass Axiom
Similar to	≡	EquivalentClass Axiom
Coordinate	⊥	DisjointClass Axiom

词义扩展器的基本思想是使用 WordNet 寻找与单词合适词义相关联的词义集合以弥补该单个词义带来的不足.比如单词 *WA* 以及它的合适词义 *WAS*,通过 WordNet 里的 *pertainym*、*derived_from_adj*、*derivationally*, *related* 关系可以得到与 *WAS* 相关联的词义集合(简记为 *WA* 的扩展词义).在匹配过程中,尽管单词的合适词义及其扩展词义不能保证为该单词找到最好的词义,但是它扩展了单词的合适词义的范围,这有助于在现存的上下文中尽可能的包括单词相关词义,并提高本体匹配过程的覆盖率.在上述例子中,单词"*published*"的合适词义是 *publish*, *bring out*, *put out*, *issue*, *release*: *prepare and issue for public distribution or sale*,其扩展词义如下:

1. *publication*, *publishing*: *the business of issuing printed matter for sale or distribution*.

2. *issue*, *publication*: *the act of issuing printed materials*.

3. *publication*, *publishing*: *the business of issuing printed matter for sale or distribution*.

4. *publisher*, *newspaper_publisher*: *the proprietor of a newspaper*.

5. *publisher*: *a person engaged in publishing periodicals or books or music*.

6. *publisher*, *publishing_house*, *publishing_firm*, *publishing_company*: *a firm in the publishing business*.

单词"*publisher*"的合适词义是 *publisher*, *publishing house*, *publishing firm*, *publishing company*: *a firm in the publishing business*.单词"*publisher*"的合适词义与单词"*published*"的扩展词义中的第 6 条词义相同,这为匹配 *publishedBy* 和 *hasPublisher* 提供了机会.

单词的合适词义以及该词义的扩展构成该单词的信息词义集(a set of informative senses).

4.2 词义关系本体的构建

在发现本体里单词的合适词义及其扩展词义之后,需要计算来自不同本体的词义之间的关系(如果存在),并根据表 1 中的转化规则,将这些关系转化为描述逻辑公理以构建词义关系本体.

例如,来自样例本体 *Ontology*101 中单词"*School*"的词义是[*Synset*:[*Offset*:32311][610][*POS*:*noun*] *Words*: *school* (*an educational institution*;)],用 *SA* 来表示;来自样例本体 *Ontology*302 中单词"*Institution*"的词义是[*Synset*:[*Offset*:42323][26][*POS*:*noun*] *Words*:*institution*, *establishment* (*an organization founded and united for a specific purpose*;)],用 *SB* 来表示.在 *WordNet* 中 *SA* 和 *SB* 之间存在上下位关系 *Hypernym*(*SB*, *SA*),即 *SB* 是 *SA* 的下位词,根据表 1 中规则,这个关系被翻译成 *SubClassAxiom* (*School*32311, *Institution*42323)公理,并被添加到词义关系本体中.

从建立词义关系本体的过程可以看到,单词信息词义集中的元素构成了词义关系本体的概念,它们之间的关系构成了词义关系本体的描述逻辑公理.词义

关系本体仅包含概念并不包含属性.在匹配样例本体过程中,产生的词义关系本体包含 907 个概念,1202 个子类公理,125 个定义公理和 168 个父类公理.

4.3　实体标记的定义

基于实体描述(实体标签和实体评论)中单词的信息词义集以及描述逻辑元素,实体词法信息标记器形式化的定义了实体标记.实体标记反映了该实体在本体中的词法信息,表达了如果不考虑来自描述逻辑的约束时,知识工程师表达的概念.

实体标记简记为 $N(E)$,其中 E 表示本体里的实体.$N(E)$是一个集合,这个集合的每个元素是由词义关系本体里的概念(或该实体描述中单词的信息词义集中元素)的交(Object Intersection)构成.实体标记的定义如下:

$$W(E)=\{e\mid e=La(E)\otimes(\sum_{m}^{i}\oplus C_n^i(Co(E)))\}$$
$$1.0\leqslant i\leqslant n$$
$$2.\ n=Co(E).size \tag{1}$$

$$IM(E)=\{s\mid s=s_1\cap s_2\cap s_2\cap s_i\cap s_{i+1}\cdots\cap s_n\}$$
$$1.\ e\in W(E)$$
$$2.\ n=e.size$$
$$3.\ W_i\in e \tag{2}$$
$$4.\ s_i=\begin{cases}special\ symbol, & \text{if } W_i\in SpWdSet\\ \in W_i.IS, & \text{otherwise}\end{cases}$$

在公式 1 中,$La(E)$是来自实体 E 的标签里所有单词的集合,$Co(E)$是来自实体 E 的评论里所有单词的集合,这两个集合里都不包括停用词(比如 *have*、*has*、*is*、*was* 等).以样例本体 *Ontology*101 中的"*Book*"概念为例,$La(101:Book)=\{Book\}$,$Co(E)=\{book, monograph, or, collection, written, text\}$.$C_n^i(A)$表示一个集合,该集合的每个元素是在 A 的 n 个元素中选 i 个得到的,$C_6^1(Co(101:book))=\{\{book\},\{monograph\},\{or\},\{collection\},\{written\},\{text\}\}$.

$\oplus$运算表示将两个集合合并为一个集合.例如,$C_6^1(Co(101:book))\oplus C_6^6(Co(101:book))=\{\{book\},\{monograph\},\{or\},\{collection\},\{written\},\{text\},\{book, monograph, or, collection, written, text\}\}$.

$\otimes$运算表示形成一个新的集合,该集合的每个元素包含 $La(E)$的所有元素和$\sum_n^i\oplus C_n^i(Co(E))$中某个集合中的所有元素,并删掉重复的单词.例如,$La(Co(101:Book))\otimes C_6^6(Co(101:book))=\{\{Book, monograph, or, collection, written, text\}\}$.

在公式 2 里,e 是公式 1 中 $W(E)$的一个元素;W_i 是 e 的一个元素;$W_i.IS$ 是单词 W_i 的信息词义集(见 4.1 节);*SpWdSet* 表示特殊的单词集(比如 *or*、*of*、*without* 等);如果 W_i 不是 SpWdSet 的元素,则 s_i 是 W_i 的信息词义集中的一个元素,否则是 SpWdSet 中的元素;$\cap$是若干词义的交.

$N(E)$中的元素有如下 4 个特点:

①该元素是 OWL 中概念交的实例(instance of OWL Object Intersection);

②该元素中的操作数包含实体标签中每个单词的信息词义集中的一个词义.

③该元素中的操作数可能包含实体评论中单词的词义集中的一个词义.

④该元素中的操作数都是词义关系本体中的概念,也就是说,该元素是词义本体中的匿名概念.

在定义实体标记过程中,不难发现,相对于实体评论,实体标签起了更重要的作用,因此实体标记的元素一定包含实体标签单词的词义.相对于实体标签,实体评论起到对实体标签的补充作用,因此实体标记的元素有可能包含一个或多个评论单词中的词义.这是因为当知识工程师构造本体时,将会为实体标签选择必要的单词,也通过实体评论进一步的描述实体.

此外,实体标签通过如下方式处理特殊单词集中的元素.例如,如果遇到词组"*reference or conference*"时,"*or*"将被翻译为,因此得到"*reference* $\cup$ *conference*";如果遇到词组"*book without report*"时,"*without*"将被翻译为$\cap$和$\neg$,因此得到"*book* $\cap\neg$ *report*".在公式(2)中,如果 s_i 是特殊单词集中的元素,需要按照特殊规则处理它们.例如,$N(101:Book)$的一个元素是$\{Book, monograph, or, collection\}$,根据公式(2),首先得到$\{book56434\ monograph534534\cap or\cap collection23244\}$,在处理特殊单词后,得到最后的结果$\{book56434\cap(monograph534534\cup collection23244)\}$.

当为属性计算实体标记时,也需要考虑属性的域(*domain*)和范围(*range*),因为它们也提供了重要的信息.例如有如下两个属性:

①*PropertyA*: < *Art* **book** *String* >

②*PropertyB*: < *Book* $\cup$ *Report* **title** *String* >

其中黑体单词是属性标签,左面表示属性的域,右面表示属性的范围.如果仅用属性标签"*book*"并不能充分表达 *PropertyA* 的词法信息.因此,当定义属性的实体标记时,不仅使用属性的标签,也使用属性域和范围里概念的标签以弥补仅使用属性标签所导致的有限性.解决此问题的过程如下:

(1)得到属性 P 的域和范围里概念的标签,分别记作 D 和 R.

(2)基于属性域和范围里的逻辑(范围里的逻辑仅针对对象属性)分别计算$\sum_n^i\oplus C_n^i(Co(D))$和$\sum_n^i\oplus C_n^i(Co(R))$.

(3)将$\sum_n^i \oplus C_n^i(Co(D))$和$\sum_n^i \oplus C_n^i(Co(R))$插入公式1里的$Co(P)$.

(4)计算公式(1)和公式(2)得到$N(P)$.

以上述的*PropertyB*为例解释定义属性实体标记的过程:

(1)得到$D=\{Book, Report\}$.如果属性的范围是数据类型,比如"*String*"、"*Char*"等,则忽略R的计算,所以*PropertyB*的R是空集.

(2)计算$\sum_n^i \oplus C_n^i(Co(Book \cup Report))$,

当$i=1$, $C_2^1(Book \cup Report)=\{\{Book\},\{Report\}\}$;

当$i=2$, $C_2^2(Book \cup Report)=\{\{Book \cup Report\}\}$;$\sum_n^i \oplus C_n^i(Co(Book \cup Report))=\{\{\},\{Book\},\{Report\},\{Book \cup Report\}\}$.

(3)计算$Co(P)=\{\{\},\{Book\},\{Report\},\{Book \cup Report\},\{title\}\}$.

(4)当定义"*Book* ∪ *Report*"的词义时,需要使用两个词义的并(*union*),一个来自单词"*Book*",另一个来自单词"*Report*",最后的词义形式是<*SenseOfBook* ∪ *SenseOfReport*>.该阶段剩下的部分与公式(1)和公式(2)所示计算方法相同.

尽管该方法某种程度上能够加强实体标记对属性词法信息的表达,但也会导致匹配的不精确.以样例本体*Ontology*101中的属性<*Academic LecturesNotes* 101:*school String*>和*Ontology*302中的属性<*Publication* 302:*notes String*>为例,根据属性实体标记的处理办法,将得到$N(302:school)$的一个元素((*academic*15216 ∪ (*lecture*8932 ∩ *Anotes*65380)) ∩ *school*46212)和$N(101:notes)$的一个元素(*publication*65022 ∩ *notes*65380),并且在词义关系本体中存在两个公理(*publication*65022 ⊇ *notes*65380)和(*notes*65380 ⊇ *academic*15216),所以根据3.4节所示方法推理出$N(101:notes) \supseteq N(302:school)$(如表2所示).

表2 类公理与匹配关系之间的转换规则

Axiom	Symbol	Symbol
SupClass	⊇	*include*
SubClass	⊆	*beInclude*
DisjointClass	⊥	*disjoint*
EquivalentClass	≡	*equivalent*

4.4 候选匹配的推理

以词义关系本体作为推理机的知识库,候选匹配生成器根据实体标记里的词法信息推理出候选匹配,为进一步的语义分析和匹配提供基础.

候选匹配(简记为MC)表达了实体间词法信息的逻辑关系,定义如下:

$$MC(A,B)=<A,B,Relation>$$

其中分别来自不同本体的实体A和B具有相同的实体类型(属性或概念);*Relation*有四种类型:包括(*include*,⊇)、被包括(*beInclude*,⊆)、相等(*equivalent*,≡)、不相交(*disjoint*,⊥).

候选匹配推理器通过两个步骤完成候选匹配推理.例如,有来自不同本体的实体标记$N(E1)$和$N(E2)$,候选匹配推理器首先使用推理机中的功能推理任意两个分别来自$N(E1)$和$N(E2)$的概念之间的关系(比如*subClass*、*supClass*、*equivalentClass*或*dijointCLass*),然后根据表2的转换规则将推理出的概念关系转化为匹配中的关系,并记录推理过程中每种关系的次数.推理过程中使用的知识库是词义关系本体,$N(E1)$和$N(E2)$的元素可以看成词义关系本体的匿名概念(如4.3节所示).

候选匹配推理器然后选择具有次数最大的关系作为$E1$和$E2$之间的候选匹配关系.如果分别来自$N(E1)$和$N(E2)$的元素之间没有任何关系,将不建立$E1$和$E2$之间的*MC*.

候选匹配推理器将产生一组候选匹配(表3出示了匹配样例本体过程中产生的部分候选匹配)为语义分析提供了基础.

表3 部分候选匹配

	101:Entity	302:Entity	Relation
1	101:Collection	302:Publication	*beIncluded*
2	101:School	302:TechReport	*include*
3	101:Book	302:Book	*equivalent*
4	101:Booklet	302:Book	*beIncluded*
5	101:author	302:author	*equivalent*
6	101:author	302:firstauhtor	*include*
7	101:howPublished	302:organization	*beIncluded*

5 实验评估

本节讨论了节2中原型系统(*OACLEI*)的评估,并探讨了基于词义组合的词法分析算法在该系统中的作用.*OACLEI*使用多种开源包用来操作本体、执行推理、比较匹配和与*WordNet*交互,包括*JENA*, *MIT Java WordNet Interface*, *Java WordNet Library API*, *Pellet API*, *OWL2.0 API*, *Protégé API*, *SKOS API*, *JOWL*和*Alignment API*等.

本节主要使用三种性能指标(准确率(*precision*)、覆盖率(*recall*)和F测度值(*F-Measure*))评估*OACLEI*系统的性能.其中F测度值反映了准确率和覆盖率的综合值(关于它们的计算,请参阅文献[5]).

5.1 实验数据

本节选择了OAEI 2009里基准测试集中书目领域的四个工业本体作为目标本体,它们来自文件夹301-

304,选择文件夹 101 里的本体作为源本体.同时,这些文件夹也提供了每次本体匹配任务时的标准匹配.通常,OAEI 2009 里的本体包括大约 37 个概念、72 个属性和 108 个公理.

5.2　评估结果

5.2.1　词法分析算法中不同组件的作用

如图 2 所示,主要有三个组件影响 OACLEI 系统匹配过程的性能:词义扩展器、实体词法信息标记器,候选匹配过滤器.为了测试各组件在 OACLEI 系统中的作用,以书目领域本体作为测试集,每次测试时去掉某个组件以检查该情况下 OACLEI 系统的性能.如图 2 所示,相对于缺少某个组件,在 OACLEI 系统中所有组件结合在一起将产生更好的性能.图 2 也说明了这些组件对 OACLEI 系统的影响程度.

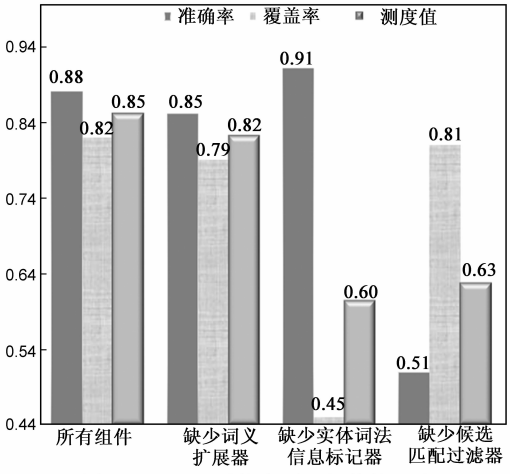

图2　OACLEI系统中不同组件的影响

在图 2 中,词义扩展器、实体词法信息标记器和候选匹配过滤器都能提高 OACLEI 系统的性能.如果缺少实体词法信息标记器,相对于完整的 OACLEI 系统,尽管准确率增加 3%(从 88%到 91%),但它的覆盖率降低了 20%(从 80%到 60%).因为如果缺少实体词法信息标记器,将破坏产生候选匹配的基础,导致较低的覆盖率,所以实体标记的定义有助于提高系统的覆盖率.

如果缺少候选匹配过滤器,准确率将减少 37%,覆盖率减少 1%,F 测度值减少 22%,因此候选匹配过滤器有助于提高系统的准确率.如果没有候选匹配过滤器,OACLEI 的覆盖率降低了 1%,然而这个组件只是删除部分候选匹配,不产生任何新的匹配.这说明一些候选匹配的存在将对系统剩余部分匹配的构造产生影响.

词义扩展器在覆盖率和准确率方面都提高了 3%.尽管该组件对系统性能影响较小,但由于它所产生的匹配(如 < 101: *date*, 302: *publishedOn*, *beIncluded* >)较少出现在其它系统的结果中.

5.2.2　与其他系统的比较

本小节通过测试书目领域的本体比较 OACLEI 和其他 7 个系统,包括 aroma、ASMOV、RiMOM、GeRoMe、Kosimap、TaxoMap 以及一种基本的匹配系统,即 edna.如图 3 和图 4 所示,OACLEI 显示出较好的性能.

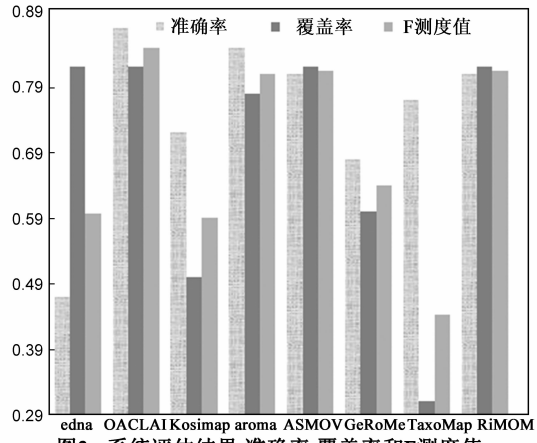

图3　系统评估结果:准确率,覆盖率和F测度值

图 3 显示了 8 个系统的准确率、覆盖率和 F 测度值.从图 3 的绿色柱图可以看到,OACLEI 在 F 测度值方面比其它系统有一定的提高:相对于 edna,提高了 25%;相对于 aroma,提高了 4%;相对于 ASMOV,提高了 4%;相对于 GeRoMe,提高了 20%;相对于 Kosimap,提高了 25%;相对于 TaxoMap,提高了 40%;相对于 RiMOM,提高了 4%.从图 3 的灰色柱图也可以看到,OACLEI 在准确率方面也有较好的值:相对于 edna,提高了 41%;相对于 aroma,提高了 38%;相对于 ASMOV,提高了 34%;相对于 GeRoMe,提高了 20%;相对于 Kosimap,提高了 16%;相对于 TaxoMap,提高了 11%;相对于 RiMOM,提高了 7%.

图 3 也出示了 aroma 有较高的准确率.因为它仅产生了大约 50 个匹配,其它系统通常产生大约 70 个的匹配;edna 的覆盖率较高,但是它的准确率较低,这说明 edna 发现了较多的匹配,但是并不十分精确.

在信息检索领域,准确率的提高往往伴随着覆盖率的降低,同时,覆盖率的提高也往往使准确率减低.图 4 出示了在准确率和覆盖率之间权衡(tradeoff)的情况.它表示在获得 n%的覆盖率时系统的准确率.在图 4 中可以看到,当覆盖率在[0%,20%]变化时 OACLEI、edna、ASMOV 和 RiMOM 有较好的准确率;当覆盖率在[20%,50%]变化时 ASMOV 和 aroma 有最好的准确率;当覆盖率在[50%,60%]变化时,OACLEI 和 Lily 有较好的准确率;当覆在率在[60%,90%]变化时,OACLEI 有较好的准确率;当覆盖率在[90%,100%]变化时,每个系统有相似的准确率.

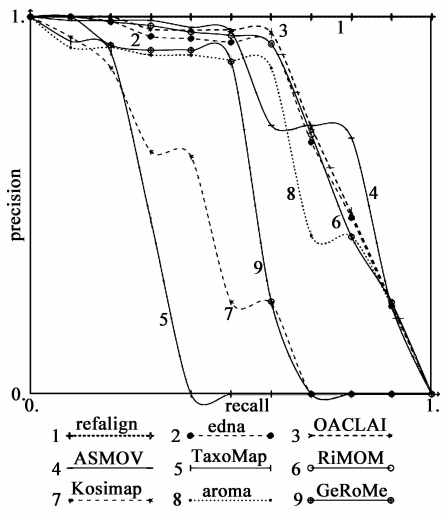

图4 系统的准确率和覆盖率之间的权衡

6 结论

基于词义组合的词法分析算法首先使用 WordNet 里特殊的词法关系,扩展单词的合适词义到一组词义,然后通过组合单词的多个词义以及语义元素(例如∩,∪),形式化的定义实体词法信息的标记.单词合适词义的扩展表示本体里的单词在目前的上下文中可能的意思,从而克服仅发现单词合适词义的缺点,这有助于发现潜在的实体匹配.实体词法信息的标记蕴含相应实体在本体里的词法信息.通过实验数据的分析,证明该算法有助于提高系统的覆盖率.

参考文献

[1] Shvaiko P, Euzenat J. Ten challenges for ontology matching [A]. Proceedings of the OTM 2008 Confederated International Conferences, CoopIS, DOA, GADA, IS, and ODBASE 2008 [C]. Monterrey, Mexico, 2008, 53(32): 300 – 313.

[2] Zhao H. Semantic matching across heterogeneous data sources [J]. Communication ACM, 2007, 50(1): 45 – 50.

[3] Kalfoglou Y, Schorlemmer M. Ontology mapping: the state of the art [J]. Knowl Eng Rev, 2003, 18(01): 1 – 31.

[4] Giunchiglia F, Shvaiko P. Semantic matching [J]. Knowl Eng Rev, 2003, 18: 265 – 280.

[5] Jean Y R, Shironoshita E P, Kabuka M R. Ontology matching with semantic verification [J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2009, 7(3): 235 – 251.

[6] Albagli S, Ben-Eliyahu-Zohary R, Shimony S. E. Markov network based ontology matching [A]. Proceeding IJCAI'09 Proceedings of the 21st International Jont Conference on Artifical Intelligence [C]. San Francisco, CA, USA, 2009.

[7] Buccella A, Cechich A, et al. Colagrossi. Building a global normalized ontology for integrating geographic data sources [J]. Computers & Geosciences, 2011, 37(7): 893 – 916.

[8] Udrea O, Getoor L, Miller R J. Leveraging data and structure in ontology integration [A]. SIGMOD 07: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data [C]. New York, NY, USA: ACM, 2007. 449 – 460.

[9] Haspelmath M. The indeterminacy of word segmentation and the nature of morphology and syntax [J]. Folia Linguistica, 2011, 45(1): 31 – 80.

[10] James N, Todorov K, Hudelot C. Combining visual and textual modalities for multimedia ontology matching [J]. Semantic Multimedia, ser. Lecture Notes in Computer Science, T. Declerck, M. Granitzer, Springer, 2011, 6725: 95 – 110.

[11] Algergawy A, Massmann S, Rahm E. A clustering-based approach for large-scale ontology matching [J]. Advances in Databases and Information Systems, ser. Lecture Notes in Computer Science, 2011, 6909: 415 – 428.

[12] Spiliopoulos V, Vouros G A, Karkaletsis V. On the discovery ofsubsumption relations for the alignment of ontologies [J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2010, 8, (1): 69 – 88.

[13] Mouselly-sergieh H, Unland R. Irom: Information retrieval-based ontology matching [J]. Semantic Multimedia, ser. Lecture Notes in Computer Science, 2011, 6725: 127 – 142.

[14] Bock J, Hettenhausen J. Discrete particle swarm optimization for ontology alignment [J]. Information Sciences, 2010, 192: 152 – 173.

[15] Tang J, Li J, Liang B, Hhuang X, et al. Using Bayesian decision for ontology mapping [J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2006, 4(04): 243 – 262.

[16] 廖建新,刘秀磊等.扩展结构包含推理算法的本体匹配[J].通信学报,2012,33(7):190 – 199.

作者简介

刘秀磊 男,1981 年生于河南,北京邮电大学博士生,主要研究方向为语义 Web、本体匹配等.
E-mail: xiuleiliu@hotmail.com

廖建新 男,1965 年生于四川,北京邮电大学教授,博士后,网络与交换技术国家重点实验室网络智能研究中心主任,主要研究方向为业务网络智能化.

朱晓民 男,1974 年生于浙江义乌,北京邮电大学副教授,博士,硕士生导师,主要研究方向为智能网、下一代业务网络、3G 核心网、协议工程等.

杨 迪 女,1985 年生于辽宁,北京邮电大学博士生,主要研究方向为 P2P 网络、搜索引擎等.

徐 童 男,1977 年生于新疆乌苏,北京邮电大学讲师,博士,目前主要从事移动互联网领域研究.