ISSN 1000-1239/CN 11-1777/TP
45(8): 1338-1345, 2008

一种新的 Web 异构语义信息搜索方法

黄 瑞[1,2] 史忠植[1]
[1](中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)
[2](中国科学院研究生院 北京 100049)
(huangr@ics.ict.ac.cn)

A New Approach to Heterogeneous Semantic Search on the Web

Huang Rui[1,2] and Shi Zhongzhi[1]
[1](*Key Laboratory of Intelligent Information Processing*, *Institute of Computing Technology*, *Chinese Academy of Sciences*, *Beijing* 100190)
[2](*Graduate University of Chinese Academy of Sciences*, *Beijing* 100049)

Abstract Relevance ranking is a key to Web search in determining how results are retrieved and ordered. As keyword-based search does not guarantee relevance in meanings, semantic search has been put forward as an attractive and promising approach. Recently several kinds of semantic information have been adopted in search respectively, such as thesauruses, ontologies and semantic markups, as well as folksonomies and social annotations. However, although to integrate more semantics would logically generate better search results, search mechanism to fully adopt different kinds of semantic information is still in absence and to be researched. To these ends, an integrated semantic search mechanism is proposed to incorporate textual information and keyword search with heterogeneous semantic information and semantic search. A statistical based measurement of semantic relevance, defined as semantic probabilities, is introduced to integrate both keywords and four kinds of semantic information including thesauruses, categories, ontologies and folksonomies. It is calculated with all textual information and semantic information, and stored in a newly proposed index structure called semantic-keyword dual index. Based on this uniform measurement, the search mechanism is developed that fully utilizes existing keyword and semantic search mechanisms to enhance heterogeneous semantic search. Experiments show that the proposed approach can effectively integrate both keyword-based information and heterogeneous semantic information in search.

Key words semantic search; semantic Web; ontology; folksonomy; social annotation

摘 要 相关排序是 Web 搜索的关键技术之一.为提高相关排序的准确性,保证搜索结果的语义相关性,语义搜索研究引入了由不同语义模型所表示的各种语义信息,如词典、语义标记、社会标注等.为了结合各类语义信息进行搜索,提出了一种新的 Web 异构语义信息搜索方法,给出了语义相关概率的定义,提出了一种基于统计的语义相关度计算方法,同时利用现有的关键词和语义搜索引擎,实现了结合关键词和异构语义信息的 Web 搜索.初步实验证明该方法可以融合关键词信息和用多种模型表示的语义信息,有效实现 Web 异构语义搜索.

关键词 语义搜索;语义 Web;本体;大众分类法;社会标注

中图法分类号 TP18

收稿日期:2007-12-19;**修回日期**:2008-04-22

基金项目:国家"九七三"重点基础研究发展规划基金项目(2007CB311004);国家"八六三"高技术研究发展计划基金项目(2006AA01Z128);国家自然科学基金项目(90604017,60435010,60775035)

相关排序是 Web 搜索的关键技术之一,其准确性很大程度上决定了搜索结果的质量.由于基于关键词的搜索无法保证语义的相关度,最近的研究将语义信息引入相关排序以提高搜索结果的语义相关度.这种基于语义信息的搜索,即所谓语义搜索,正逐渐成为研究的热点[1-4].

语义搜索研究引入不同的语义模型及以其表示的各种语义信息,如潜在语义索引(LSI)模型[5]、词典、分类法、语义 Web[6] 中的本体[7] 和语义标记、Web 2.0[8] 中的大众分类法和社会语义信息等.

图 1 为关键词和异构语义信息例,包括 LSI 中的标引项、词典中的词或词组、ODP 分类①、基于本体的语义标记和 del.icio.us② 中的社会标注[9].

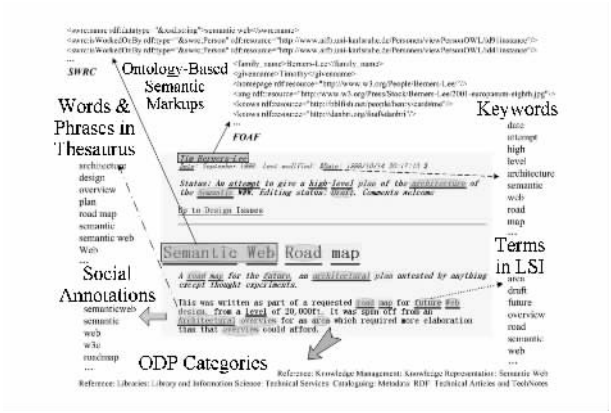

Fig. 1　Example heterogeneous semantic information.

图 1　Web 中的异构语义信息例

这些异构语义模型对语义搜索提出了新的挑战.由于不同语义模型所表示的语义信息都从一定程度或一定侧面描述了内容的一些语义,因此在搜索中更多地结合这些语义信息可以更深入地理解内容的语义,更准确地对检索结果进行语义相关排序,从而在语义层面上提高查全率和查准率.然而据我们所知,目前还难以将关键词与异构语义信息结合进行搜索.因此,亟需研究快速有效的 Web 信息搜索方法,充分利用多种异构语义信息提高搜索质量.

本文提出一种新的异构语义信息搜索方法,针对关键词和异构语义信息定义了语义相关概率,提出一种结合关键词和异构语义的查询扩展算法,和一种统计语义相关度算法;利用现有的关键词和语义搜索引擎,实现结合关键词和异构语义信息的 Web 搜索,并通过实验验证了其可行性和有效性.

1　相关工作

现有的语义搜索方法可以大致分为 4 类:

1) 基于潜在语义索引(latent semantic indexing,LSI)模型的方法[5].通过对训练文档集中的词-文档频率矩阵进行变换,获得潜在语义索引(矩阵),数值化词/词组与文档之间的语义相关性,用于在信息搜索时进行语义相似度计算.该方法无须领域相关的先验知识,而是从训练文档集中学习获得,且无须手工标注语义,因此可用于搜索大量混杂的文档集.然而,LSI 模型中的语义知识无法供人理解,模型的性能很大程度上依赖于训练文档集的质量,且难以支持高层的基于语义的应用.

2) 基于词典(thesaurus)的方法[10-12].预先定义词典,给出词、词组及其具体的含义和一些简单的相互关系(如同义词、反义词),在搜索时到词典中进行查询,获得相关语义,用于词义消歧[10]、查询扩展[11]、计算语义相似度[12]等,实现语义搜索.词典包括较为广泛的领域和主题,所表示的语义比较清晰、通用,容易被理解,且无须手工标注语义,因此可以用于大规模、混杂但拥有相对稳定的背景知识的信息搜索,并支持进一步基于语义的应用.然而,词典通常需要手工创建,不易修改,通常仅限于表征通用的浅层语义,难以深入、准确地表征复杂的语义.

3) 基于本体(ontology)和语义标记(markup)的方法[1-3,13-15].预先定义领域本体,形式化地描述领域公认的概念和概念间的关系,为领域知识的表示提供术语,在编写信息(或进行语义标注)时对语义进行基于本体的明确标注;搜索时不仅可以通过明确的标注和相应的本体理解信息的语义,还可以进行基于语义的逻辑推理[1,13-14];最新的研究将语义推理与关键词检索结合[2-3,15],提高搜索性能.本体描述了清晰、明确、易于理解的语义,支持复杂的逻辑推理和高层基于语义的应用.然而,大部分本体由手工创建与维护,基于本体的语义标记也需要手工参与,因此,目前 Web 上海量、动态、混杂的内容难以用本体完全标注.此外,目前基于本体的推理机制因其复杂性也难以适应大规模、动态的推理.

4) 基于大众分类法(folksonomy)和社会标注的方法[4,9,16-17].社会标注是指由大量普通 Web 用

① 开放目录项目 Open Directory Project, http://dmoz.org/

② http://del.icio.us

户手工对信息添加的标签(如图片的标注、网页的书签等).由大量社会标注组成的平面非等级的标签分类被称为大众分类法,以其代替传统的关键词[4,16]或与关键词一起[9,17]理解内容的语义并搜索语义相关的信息.这类方法适合于有大量用户参与的开放、动态、混杂的 Web 环境,如 Web 2.0.然而,社会标注由用户随意标注,通常比较模糊且简短不规范,尽管可以被用户理解却难以对其语义进行管理和应用,在语义搜索中需要预先进行词处理(如组合词切分)、聚类或统计分析后才能获得比较满意的结果.

虽然有时不明确区分本体和词典,但本文中,基于词典的方法在编写信息时无须对照词典进行标注,仅在搜索时查询词典获取语义.基于本体和语义标记的方法中,本体基于描述逻辑,要在其基础上给出明确、规范的语义标记,搜索时不仅能获取语义而且能支持推理.分类法和基于分类法的搜索中,分类要预先定义,且对信息要根据分类法手工进行分类,类似第 3)类,因此不单独分析.

本文目标是结合关键词和异构的语义信息进行搜索.文献[2]明确提出应将语义推理与关键词搜索相结合,并通过搜索结果融合和相关反馈实现这种结合.文献[3]采用模糊描述逻辑实现推理与检索的结合.本文采用基于统计的方法,充分结合关键词和词典、本体等,实现查询扩展、语义元搜索和语义相关排序.本文提出的统计语义相关度算法基于 LSI 模型的思想,并受到文献[9]的启发,他们利用社会标注的统计信息计算用户的社会关系和文档的语义相似度.文献[17]结合关键词与企业网内的用户标注共同创建索引进行搜索;本文针对开放 Web 中的多种语义模型实现基于异构语义信息的搜索.

2 基于统计的异构语义信息搜索方法

2.1 概 述

本文提出的基于统计的异构语义信息搜索框架如图 2 所示.文档库中包括关键词和语义信息,用于语义知识学习(见第 2.3 节)和语义索引(见第 2.4 节).知识库中的其他知识,如词典和本体,则由知识工程师导入和维护.当服务器端建好初始知识库和索引库后,系统就可以在客户端运行了.当用户输入查询或相关反馈时,首先进行语义查询扩展(见第 2.5 节),并将扩展后的查询输入元搜索引擎进行搜索.然后结合搜索结果和知识库中的知识,融合所有关键词和语义搜索结果,返回给用户(见第 2.6 节).最后,将不在文档库中且与原查询的相关度大于一定阈值的新文档加入文档库;当这类新文档的数目大于设定阈值时,重复进行语义索引和语义知识学习,更新索引库和知识库.

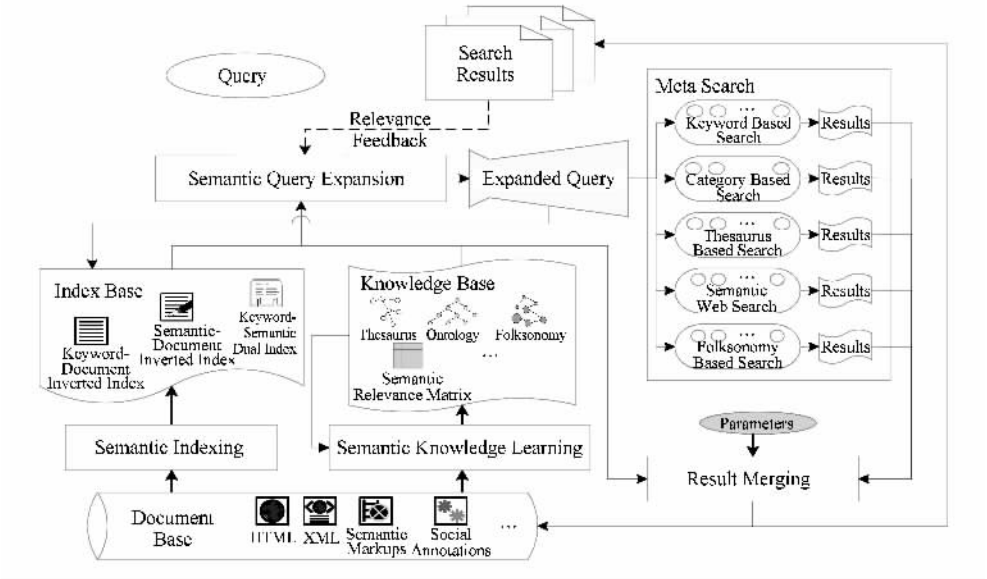

Fig. 2 Statistical semantic search framework.

图 2 基于统计的异构语义信息搜索框架

2.2 定 义

本文在传统的统计检索模型[18]基础上,针对语义空间中的关键词和异构语义,给出了语义相关概率的定义,利用概率统计方法处理异构的语义信息.

语义空间 $SA=(sa_1,\cdots,sa_p)$ 是所有(p 个)语义标注对的集合.一个语义标注对 $sa=[\boldsymbol{T},\boldsymbol{S}]$ 表示

语义空间中的一个关键词组-语义单元组对,$\boldsymbol{T}$ 为该语义标注对中所有关键词的列表,$\boldsymbol{S}$ 为该语义标注对中所有语义单元的列表. 语义单元指可独立表示清晰语义的最小单位,如词典中的一个词、本体中的一个概念、大众分类法中的一个社会标注. $\forall t_i \in \boldsymbol{T}$ 为关键词列表 $\boldsymbol{T}$ 中的第 i 个关键词,t_i 出现在语义标注对 $sa=[\boldsymbol{T},\boldsymbol{S}]$ 中记作 $t_i \vdash_{sa}^{T}$. $\forall s_j \in \boldsymbol{S}$ 表示语义单元列表 S 中的第 j 个语义单元,s_j 出现在语义标注对 $sa=[\boldsymbol{T},\boldsymbol{S}]$ 中记作 $s_i \vdash_{sa}^{S}$. 若 $t_i \vdash_{sa}^{T} \wedge s_j \vdash_{sa}^{S}$,则称 t_i 和 s_j 共现于语义标注对 $sa=[\boldsymbol{T},\boldsymbol{S}]$ 中,记作$\langle t_i, s_j\rangle \vdash_{sa}$. 图 1 对应的语义空间片断如图 3 所示:

```
([("Semantic","Web","Road","map",…),
  ("Reference: Knowledge Management…","Reference:…")],
[("semantic","Web"),
  ("〈swrc:name〉","〈swrc:isWorkedOnBy〉",…)],
[("Semantic","Web","Road","map",…),
  ("semanticweb","semantic","web","w3c",…)],
…)
```

Fig. 3　Part of the semantic space constructed from Fig. 1.

图 3　从图 1 构造的语义空间片断

在此基础上定义关键词或语义单元在语义空间中的语义出现概率、语义共现概率和语义条件概率.

定义 1. 语义出现概率. $\forall (t_i \vdash_{sa}^{T}) \wedge (sa \in SA)$,$P(t_i) \in [0,1]$表示关键词 t_i 在语义空间 SA 中的语义出现概率. $\forall (s_j \vdash_{sa}^{S}) \wedge (sa \in SA)$,$P(s_j) \in [0,1]$表示语义单元 s_j 在语义空间 SA 中的语义出现概率.

语义出现概率表示关键词或语义单元在整个语义空间中出现的概率.

定义 2. 语义共现概率. $\forall (\langle t_i, s_j\rangle \vdash_{sa}) \wedge (sa \in SA)$,$P(\langle t_i, s_j\rangle) \in [0,1]$表示关键词 t_i 和语义单元 s_j 在语义空间 SA 中的语义共现概率.

语义共现概率表示关键词和语义单元在整个语义空间中共同出现的概率,即在整个语义空间中,用该语义单元标注该关键词的概率.

定义 3. 语义条件概率. $P(t_i/s_j) = P(\langle t_i, s_j\rangle)/P(s_j) \in [0,1]$表示关键词 t_i 相对给定语义单元 s_j 在语义空间 SA 中的语义条件概率. $P(s_j/t_i) = P(\langle t_i, s_j\rangle)/P(t_i) \in [0,1]$表示语义单元 s_j 相对关键词 t_i 在语义空间 SA 中的语义条件概率.

关键词相对于语义单元的语义条件概率表示在整个语义空间中,在该语义单元出现的条件下,该关键词出现的条件概率. 语义单元相对于关键词的语义条件概率表示在整个语义空间中,在该关键词出现的条件下,该语义单元出现的条件概率.

2.3　语义相关性知识学习

语义知识库中除了词典、本体、分类法和大众分类法,还包括学习获得的语义相关性矩阵. 算法 1 给出了语义相关性矩阵的学习算法. 利用整个语义空间中的关键词和语义单元,基于 LSI[5] 模型的思想,计算所有关键词和语义单元的潜在语义相关性.

算法 1. 语义知识学习算法(semantic knowledge learning).

输入:语义空间 SA;

输出:语义相关性矩阵 $\boldsymbol{M}$.

Begin

① 从 SA 中统计初始的语义共现概率矩阵 $\boldsymbol{M}_0$: $\boldsymbol{M}_0[i][j]=\boldsymbol{P}(\langle \boldsymbol{T}[i], \boldsymbol{S}[j]\rangle)$;

② 奇异值分解 $\boldsymbol{M}_0=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$,其中,$\boldsymbol{U}$ 和 $\boldsymbol{V}$ 为下三角矩阵,$\boldsymbol{\Sigma}=\mathrm{diag}(\sigma_1,\cdots,\sigma_r)$为对角线矩阵;

③ 维度约减 $\boldsymbol{\Sigma}'=\mathrm{diag}(\sigma_1,\cdots,\sigma_{r'},0_{r'+1},\cdots,0_r)$;

④ 构造关键词-语义单元相关矩阵 $\boldsymbol{M}_{\mathrm{TS}}=\boldsymbol{U}\boldsymbol{\Sigma}'\boldsymbol{V}^{\mathrm{T}}$,计算语义单元-关键词相关矩阵 $\boldsymbol{M}_{\mathrm{ST}}=\boldsymbol{M}_{\mathrm{TS}}{}^{\mathrm{T}}$;

⑤ 计算关键词-关键词相关矩阵 $\boldsymbol{M}_{\mathrm{TT}}=\boldsymbol{M}_{\mathrm{TS}}\boldsymbol{M}_{\mathrm{ST}}$和语义单元-语义单元相关矩阵 $\boldsymbol{M}_{\mathrm{SS}}=\boldsymbol{M}_{\mathrm{ST}}\boldsymbol{M}_{\mathrm{TS}}$;

⑥ 构造语义相关性矩阵 $\boldsymbol{M}=\begin{bmatrix}\boldsymbol{M}_{\mathrm{TT}} & \boldsymbol{M}_{\mathrm{TS}}\\ \boldsymbol{M}_{\mathrm{ST}} & \boldsymbol{M}_{\mathrm{SS}}\end{bmatrix}$

End

首先,计算语义空间 SA 中关键词 t_i 和语义单元 s_j 的语义共现概率:

$$P(\langle t_i, s_j\rangle) = \Big(\sum_{sa=[\boldsymbol{T},\boldsymbol{S}]\in \boldsymbol{SA}, t_i \vdash_{sa}^{T} \wedge s_j \vdash_{sa}^{S}} \times \Big(\ln\frac{\|SA\| - df(t_i) + 0.5}{df(t_i) + 0.5} \times \ln\frac{\|SA\| - df(s_j) + 0.5}{df(s_j) + 0.5} \times \frac{1+\ln(1+\ln tf(t_i,\boldsymbol{T}))}{(1-\gamma)+\gamma\frac{\|\boldsymbol{T}\|}{avTl}} \times \frac{1+\ln(1+\ln tf(s_j,\boldsymbol{S}))}{(1-\delta)+\delta\frac{\|\boldsymbol{S}\|}{avSl}}\Big)\Big/\|SA\|,$$

其中,$df(t_i)$和 $tf(t_i,\boldsymbol{T})$分别表示关键词 t_i 在语义空间 SA 中的文档频率和在一个语义标注对 $sa=[\boldsymbol{T},\boldsymbol{S}]$中的关键词列表 $\boldsymbol{T}$ 中的词频;$df(s_j)$和 $tf(s_j,\boldsymbol{S})$分别表示语义单元 s_j 在语义空间 SA 中的文档频率和在一个语义标注对 $sa=[\boldsymbol{T},\boldsymbol{S}]$中的语义单元列表 $\boldsymbol{S}$ 中的词频;$avTl$ 表示平均关键词列表长度,$avSl$ 表示平均语义标注列表长度,γ 和 δ 为长度归整参数. 然后,对语义共现概率矩阵进行奇异值分解(singular value decomposition,SVD),分解为一个

下三角矩阵 $\boldsymbol{U}$、一个对角线矩阵 $\boldsymbol{\Sigma}$ 和一个上三角矩阵 $\boldsymbol{V}^{\mathrm{T}}$ 的乘积. 第 3 步,将 r 维对角线矩阵 $\boldsymbol{\Sigma}$ 降维成 r'维对角线矩阵 $\boldsymbol{\Sigma}'$,降维后的矩阵 $\boldsymbol{\Sigma}'$分别与下三角矩阵 $\boldsymbol{U}$ 和上三角矩阵 $\boldsymbol{V}^{\mathrm{T}}$ 左、右乘,得到关键词-语义单元相关矩阵 $\boldsymbol{M}_{\mathrm{TS}}$. 该矩阵基于原共现概率矩阵,将其从高维语义空间投影到较低维语义空间,表示所有关键词和语义单元之间的潜在语义相关性. 最后,从关键词-语义单元相关矩阵 $\boldsymbol{M}_{\mathrm{TS}}$经过变换和计算获得语义相关性矩阵知识,用于在查询扩展时计算推理结果的语义相关性以及在搜索结果融合时计算文档的语义相关排序.

2.4 语义索引

本文中的索引包括 3 类:第 1 类是关键词-文档索引,和传统的基于关键词的搜索引擎类似;第 2 类是语义-文档索引,即以语义单元代替关键词-文档索引中的关键词;第 3 类是本文提出的关键词-语义双向索引,将相关的关键词和语义单元相互链接,以便高效地在关键词和语义单元之间相互检索.

关键词-语义双向索引由关键词-语义单元倒排索引和语义单元-关键词倒排索引组成,见图 4:

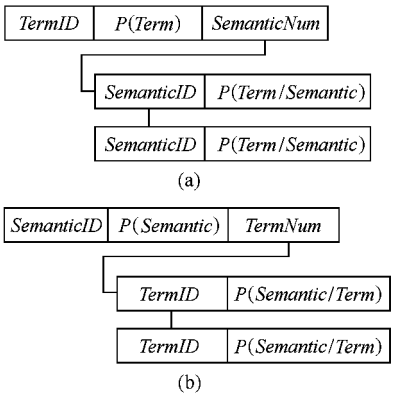

Fig. 4 Keyword-semantic dual index. (a) Semantic-term inverted index and (b) Term-semantic inverted index.

图 4 关键词-语义双向索引. (a) 关键词-语义单元倒排索引;(b) 语义单元-关键词倒排索引

关键词-语义单元倒排索引按关键词的唯一标识"*TermID*"排序. "$P(Term)$"存储该关键词 Term 在语义空间中的语义出现概率:

$$P(t_i)=\Big(\sum_{sa=[\boldsymbol{T},\boldsymbol{S}]\in SA,\,t_i\vdash_{sa}^{\boldsymbol{T}}}\Big(\ln\frac{\|SA\|-df(t_i)+0.5}{df(t_i)+0.5}\times\frac{1+\ln(1+\ln tf(t_i,\boldsymbol{T}))}{(1-\gamma)+\gamma\dfrac{\|\boldsymbol{T}\|}{avTl}}\Big)\Big)\Big/\|SA\|.$$

"*SemanticNum*"存储相关语义单元的数目,即其后"*SemanticID*"链接长度;"$P(Term/Semantic)$"表示给定相应语义单元 *Semantic*,该关键词 *Term* 的语义条件概率 $P(t_i/s_j)=P(\langle t_i,s_j\rangle)/P(s_j)$,其中的语义共现概率取第 2.3 节中经过 LSI 变换后的值. 语义单元-关键词倒排索引表的结构类似,其中

$$P(s_j)=\Big(\sum_{sa=[\boldsymbol{T},\boldsymbol{S}]\in SA,\,s_j\vdash_{sa}^{\boldsymbol{S}}}\Big(\ln\frac{\|SA\|-df(s_j)+0.5}{df(s_j)+0.5}\times\frac{1+\ln(1+\ln tf(s_j,\boldsymbol{S}))}{(1-\delta)+\delta\dfrac{\|S\|}{avSl}}\Big)\Big)\Big/\|SA\|.$$

其中参数的含义与第 2.3 节相同.

2.5 基于统计的语义查询扩展

本文方法中的查询扩展同时利用关键词和语义单元,分 3 步实现:基于语义相关概率的相关度计算、基于本体和逻辑的语义推理以及结合关键词和语义单元的查询扩展. 算法 2 描述了整个语义查询扩展算法,包括步骤 1 至步骤 3 的详细算法.

算法 2. 语义查询扩展 Semantic Query Expansion.

输入:① 已解析查询 Q ($Type$:类型,关键词或语义单元,ID:编号,W:关键词/语义单元在查询语句中的权重);

② 关键词-语义双向索引 I_{TS},I_{ST};

③ 语义知识库 KB.

输出:扩展后的查询 Q' ($Type$,ID,CP:语义条件概率).

参数:① $\theta\in[0,1]$:语义条件概率的相关度阈值;② $\kappa\in[0,1)$:语义条件概率的传递衰减因子.

中间结果:① 关键词列表 T (ID,CP);

② 语义单元列表 S (ID,CP).

Begin

步骤 1. (T,S)←Semantic computing(Q,I_{TS},I_{ST});

步骤 2. (S)←Semantic inference(KB,S);

步骤 3. (Q')←Query expansion (T,S).

End

步骤 1 在关键词-语义双向索引中找出所有与初始查询的相关度满足条件的关键词和语义单元. 循环调用子函数 $addList$,由相关度阈值 θ 和传递衰减因子 κ 控制相关关键词和语义单元在双向索引中的扩展,筛选所有满足条件的关键词和语义单元.

步骤 1. (T,S)←Semantic computing (Q,I_{TS},I_{ST}).

Begin

For all$+t\in Q$ do /* 结果中包括的关键词 */

$addList(t.ID,t.W,1,\theta,\kappa,\boldsymbol{T},\boldsymbol{S},I_{\mathrm{TS}},I_{\mathrm{ST}})$

End For
For all $-t \in Q$ do /* 结果中不包括的关键词 */
$addList(t.ID, t.W, -1, \theta, \kappa, \boldsymbol{T}, \boldsymbol{S}, I_{TS}, I_{ST})$
End For
For all $+s \in Q$ do /* 结果中包括的语义单元 */
$addList(s.ID, s.W, 1, \theta, \kappa, S, T, I_{ST}, I_{TS})$
End For
For all $-s \in Q$ do /* 结果中不包括的语义单元 */
$addListSSS(s.ID, s.W\ -1, \theta, \kappa, \boldsymbol{S}, \boldsymbol{T}, I_{ST}, I_{TS})$
End For
End

子函数 *addList*(int *id*, float *cp*, int *il*, float *th*, *float df*, List L_1, List L_2, Index I_1, Index I_2)

Begin
$L_1.add(id, il \times cp)$ /* 当 $add(id, x)$时,若有(id, y),则 $y \leftarrow y + x$ */
For $i=0$ to $I_1[id].LinkNum$ do
If $(df \times I_1[id].Link[i].CP) \geqslant th$ then
$addList(I_1[id].Link[i].ID, (df \times I_1[id].Link[i].CP), il, th, (\kappa \times df), L_2, L_1, I_2, I_1)$
End If
End For
End

步骤 2 利用语义知识库中的知识查询或推理相关的语义单元,并利用语义相关性矩阵计算推理结果与原查询的相关性. 语义推理采用现有的基于本体的查询或推理方法[18-20],对查询扩展后的语义单元列表中的每个语义单元 $s_i \in \boldsymbol{S}$,由它推理所得的每一个相关语义单元 $ns_j \in KB$,如果 ns_j 与原语义单元 s_i 的相关性大于预先设定的阈值,就认为该推理所得的语义单元 ns_j 与原查询相关,将其加入 $\boldsymbol{S}$.

步骤 2. (S) ←Semantic inference(KB,S).
Begin
For all 语义单元 $s_i \in \boldsymbol{S}$ do
For all 查询/推理所得的 ns_j in KB do
If ($si.CP \times \boldsymbol{M}_{SS}[ns_j][s_i] \geqslant$阈值) then
$S.add(ns_j.ID, s_i.CP \times \boldsymbol{M}_{SS}[ns_j][s_i])$
End If
End For
End For
End

最后,步骤 3 利用所有计算和推理所得的相关关键词和语义单元对原查询进行扩展.

步骤 3. (Q') ←Query expansion(T,S).
Begin
$maxCP \leftarrow \max(\max(t_i.CP \mid t_i \in T), \max(s_j.CP \mid s_j \in S))$
For all term $t \in T$ do
$Q'.add(\text{'}Term\text{'}, t.ID, t.CP/maxCP)$
End For
For all semantic unit $s \in S$ do
$Q'.add(\text{'}Semantic\text{'}, s.ID, s.CP/maxCP)$
End For
End

不考虑推理的时间复杂度时,算法 2 的时间复杂度为:$O(\max(\max(SN_1, SN_2, \cdots), \max(TN_1, TN_2, \cdots)))$,其中 SN_i 表示第 i 个关键词的"*SemanticNum*",TN_j 表示第 j 个语义单元的"*TermNum*",即关键词-语义单元双向索引中,与关键词相关的语义单元总数和与语义单元相关的关键词总数的最大值,由于该值通常较小,算法 2 的时间复杂度通常等于 $O(1)$.

2.6　搜索结果融合

搜索结果融合结合统计语义相关度和推理语义相关度对所有结果进行相关排序.

设 $\|\boldsymbol{T}\| = m$ 为 SA 中所有关键词的数目,$\|\boldsymbol{S}\| = n$ 为 SA 中所有语义单元的数目,语义扩展后的查询为 $\boldsymbol{q} = [\boldsymbol{q}_T\ \boldsymbol{q}_S]$,搜索结果中的文档可表示为 $\boldsymbol{d} = [\boldsymbol{d}_T\ \boldsymbol{d}_S]^T$,则 $\boldsymbol{d}$ 与 $\boldsymbol{q}$ 的统计语义相关度:

$$SS(\boldsymbol{q}, \boldsymbol{d}) = \boldsymbol{q}\boldsymbol{M}\boldsymbol{d} = [\boldsymbol{q}_T\ \boldsymbol{q}_S] \begin{bmatrix} \boldsymbol{M}_{TT} & \boldsymbol{M}_{TS} \\ \boldsymbol{M}_{ST} & \boldsymbol{M}_{SS} \end{bmatrix} \begin{bmatrix} \boldsymbol{d}_T \\ \boldsymbol{d}_S \end{bmatrix}.$$

基于语义推理所得的语义单元,可以搜索更多的语义相关结果文档,它们与原查询由推理而得的语义相关度可表示为 $IS(\boldsymbol{q}, \boldsymbol{d})$[19,21].

统计语义相关度和推理语义相关度通过参数 α 进行融合,融合后的语义相关度为

$$Sim(\boldsymbol{q}, \boldsymbol{d}) = \alpha \times SS(\boldsymbol{q}, \boldsymbol{d}) + (1 - \alpha) \times IS(\boldsymbol{q}, \boldsymbol{d})$$

不考虑推理语义相关度计算的复杂度时,语义相关度计算的复杂度为 $O((m+n)^2)$.

3　实验研究

3.1　实验数据

本文利用 Web 中的异构语义信息进行实验,包

括含 147249 个串的 WordNet 词典①，含 730416 个类和 4792967 个链接的 ODP 数据②，从 Swoogle③缓存中抽取的 10429951 个 RDF 三元组④和从 Web 中下载的 347 个本体，以及我们 2007 年 5 月抓取的含 459143 个社会语义标注的 del.icio.us 数据，目前共包括 13849511 个关键词和 97390 个语义单元，选择经验参数 $\gamma=\delta=0.3$，$\theta=0.4$，$\kappa=0.9$，$\alpha=0.7$.

3.2 初步结果

实验在多主体平台 MAGE[22] 上进行，图 5 为系统后台界面. 其中，“Main-Container”控制 MAGE，“Container-1”包括本文框架中的模块，“Container-2”到“Container-5”中目前包括 Google、ODP 搜索、Swoogle 和 del.icio.us 四类搜索引擎，

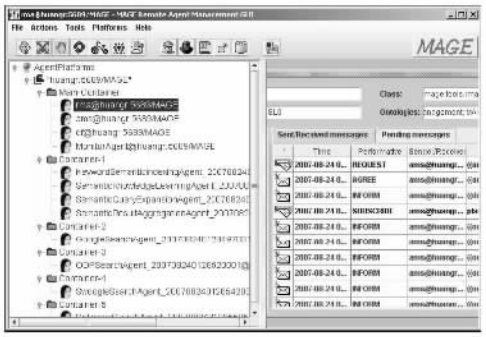

Fig. 5 System server based on MAGE.

图 5 基于 MAGE 的系统后台界面

检索时由于需要进行查询扩展、元搜索和搜索结果融合，会降低搜索的效率，根据前文中的时间复杂度分析，且由于元搜索中需等待各搜索引擎的检索结果并对网页进行分析，因此，影响效率的关键首先是元搜索的效率，其次是关键词和语义单元的总数.

图 6 给出了查询“semantic web”的搜索结果例. 系统给出了 10 个最相关的关键词和 10 个最相关的语义单元，其中最相关的关键词包括 web，semantic，rdf，ontology，xml，w3c 等，最相关的语义单元包括 *semantic*，*rdf*，*web*，*category*：*topic*，*ontology*，*srwc*：*isworkedonby*，*rdfs*：*comment* 等. 虽然实验中仅包含了部分 Web 信息，但本文方法还是给出了较合理的相关语义. 因此，本文提出的搜索引擎可以利用多种现有搜索引擎，结合关键词和语义单元进行搜索.

Fig. 6 Example search results.

图 6 搜索结果例

4 总 结

本文研究利用 Web 中异构的语义信息进行搜索的方法，提出了一种基于统计的方法，同时利用 Web 中的关键词和异构语义进行搜索. 初步实验证明本文方法可融合关键词信息和用多种模型表示的语义信息实现 Web 异构语义搜索. 下一步我们将扩大训练文档库的规模，并逐步完善系统，投入实用.

参 考 文 献

[1] Guha R, Mccool R, Miller E. Semantic search [C] //Proc of the 12th Int'l Conf on World Wide Web (WWW'03). New York: ACM, 2003

[2] Mayfield J, Finin T. Information retrieval on the semantic web: Integrating inference and retrieval [C] //Proc of the SIGIR Workshop on the Semantic Web. New York: ACM, 2003

[3] Zhang L, Yu Y, *et al*. An enhanced model for searching in semantic portals [C] //Proc of the 14th Int'l Conf on World Wide Web (WWW'05). New York: ACM, 2005

[4] Bao S, Wu X, *et al*. Optimizing Web search using social annotations [C] //Proc of the 16th Int'l Conf on World Wide Web (WWW'07). New York: ACM, 2007

[5] Furnas G W, Deerwester S, *et al*. Information retrieval using a singular value decomposition model of latent semantic structure [C] //Proc of the 11th Annual Int'l ACM SIGIR Conf on Research and Development in Information Retrieval (SIGIR'88). New York: ACM, 1988

[6] Berners-Lee T, Hendler J, Lassila O. The semantic Web [J]. Scientific American, 2001, 284(5): 34-43

[7] Studer R, Benjamins V R, Fensel D. Knowledge engineering: Principles and methods [J]. Data and Knowledge Engineering, 1998, 25(1-2): 161-197

① http://wordnet.princeton.edu/obtain

② http://rdf.dmoz.org/rdf/

③ http://swoogle.umbc.edu/

④ http://ebiquity.umbc.edu/resource/html/id/126/10M-RDF-triples

[8] O'Reilly T. What is web 2.0: Design patterns and business models for the next generation of software [OL]. (2005-09) [2007-12]. http://www.oreilly.com/

[9] Wu X, Zhang L, Yu Y. Exploring social annotations for the semantic Web [C] //Proc of the 15th Int'l Conf on World Wide Web (WWW'06). New York: ACM, 2006

[10] Voorhees E M. Using wordnet to disambiguate word senses for text retrieval [C] //Proc of the 16th Annual Int'l ACM SIGIR Conf on Research and Development in Information Retrieval (SIGIR'93). New York: ACM, 1993

[11] Voorhees E M. Query expansion using lexical semantic relations [C] //Proc of the 16th Annual Int'l ACM SIGIR Conf on Research and Development in Information Retrieval (SIGIR'94). New York: Springer, 1994

[12] Tollari S, Glotin H, Maitre J L. Enhancement of textual images classification using segmented visual contents for image search engine [J]. Multimedia Tools and Applications, 2005, 25(3): 405-417

[13] Cohen S, Mamou J, *et al*. Xsearch: A semantic search engine for xml [C] //Proc of the 29th Int'l Conf on Very Large Data Bases (VLDB'03). San Fransisco: Morgan Kaufmann, 2003

[14] Ding L, Finin T, *et al*. Search on the semantic Web [J]. IEEE Computer, 2005, 10(38): 62-69

[15] Rocha C, Schwabe D, de Aragao M P. A hybrid approach for searching in the semantic web [C] //Proc of the 13th Int'l Conf on World Wide Web (WWW'04). New York: Springer, 2004

[16] Hotho A, Jäschke R, *et al*. Information retrieval in folksonomies: Search and ranking [C] //Proc of the 3rd European Semantic Web Conf. Berlin: Springer, 2006

[17] Dmitriev D A, Eiron N, *et al*. Using annotations in enterprise search [C] //Proc of the 15th Int'l Conf on World Wide Web (WWW'06). New York: ACM, 2006

[18] Ding Guodong, Bai Shuo, Wang Bin. A survey of statistical language modeling for text retrieval [J]. Journal of Computer Research and Development, 2006, 43(5): 769-776 (in Chinese)
(丁国栋, 白硕, 王斌. 文本检索的统计语言建模方法综述 [J]. 计算机研究与发展, 2006, 43(5): 769-776)

[19] Maguitman A G, Menczer F, *et al*. Algorithmic detection of semantic similarity [C] //Proc of the 14th Int'l Conf on World Wide Web (WWW'05). New York: ACM, 2005.

[20] Shi Zhongzhi, Dong Mingkai, Jiang Yuncheng, *et al*. Logical foundations of semantic Web [J]. Science in China, Ser. E, Information Sciences, 2004, 34(10): 1123-1138 (in Chinese)
(史忠植, 董明楷, 蒋运承, 等. 语义 Web 的逻辑基础[J]. 中国科学(E辑): 信息科学, 2004, 34(10): 1123-1138)

[21] Stojanovic N, Struder R, Stojanovic L. An approach for the ranking of query results in the semantic Web [C] //Proc of the 2nd Int'l Semantic Web Conference (ISWC'03). Berlin: Springer, 2003

[22] Shi Z, Zhang H, *et al*. Mage: An agent-oriented programming environment [C] //Proc of the 3rd IEEE Int'l Conf on Cognitive Informatics. Los Alamitos, CA: IEEE Computer Society Press, 2004

Huang Rui, born in 1981. Ph. D. candidate in the Key Lab of Intelligent Information Processing, the Institute of Computing Technology (ICT), the Chinese Academy of Sciences (CAS). Her main research interests include semantic Web, information retrieval, and data mining.

黄　瑞,1981 年生,博士研究生,主要研究方向为语义 Web、信息检索、数据挖掘.

Shi Zhongzhi, born in 1941. Professor and Ph. D. supervisor at the ICT, CAS, leading the Key Lab of Intelligent Information Processing. His main research interests include intelligence science, multi-agent systems, semantic Web, machine learning and neural computing. He is a senior member of IEEE, and a member of AAAI and ACM.

史忠植,1941 年生,研究员,博士生导师,主要研究方向为智能科学、多主体系统、语义 Web、机器学习和神经计算等.

Research Background

Nowadays, search engines are heavily relied on to retrieve information on the Web. Relevance ranking is key to Web search paradigms, according to which potentially related Web documents are retrieved and ordered. As keyword-based Web search does not guarantee relevance in meanings, semantic search has recently attracted enormous research focuses. Heterogeneous semantic models have been introduced and adopted in search respectively, yet no current search mechanism fully utilizes different kinds of semantic information to enhance Web search. This paper explores four kinds of semantic knowledge to improve keyword-based Web search, including thesauruses, categories, ontologies, and folksonomies. A statistical based measurement of semantic relevance, defined as semantic probabilities, is introduced to integrate heterogeneous semantic information. The search mechanism is developed that fully utilizes both keyword and heterogeneous semantic information to enhance Web search. Our work is supported by the 973 National Basic Research Programme (2007CB311004), the 863 National High-Tech Program (2006AA01Z128), and the National Natural Science Foundation of China (90604017, 60435010, and 60775035).