图书情报工作
第 59 卷 第 4 期 2015 年 2 月

关联数据驱动的查询扩展技术研究

田野[1] 杨眉[1] 祝忠明[2] 张静蓓[3]
[1]上海交通大学图书馆 上海 200240 [2]中国科学院兰州文献情报中心 兰州 730070
[3]上海外国语大学图书馆 上海 201620

摘要: [目的/意义] 针对当前查询扩展技术面临的瓶颈,提出一种关联数据驱动的查询扩展方法,改善检索系统的查全率、查准率。 [方法/过程] 将扩散激活理论应用到关联数据集中,使得在输入查询词搜索潜在语义实体时,对提取的查询词的语义特征在知识库中进行有特定机制的扩散和激活,最后对这些语义关联的候补概念进行收集,并利用推理机制进行筛选,得到更优的概念集。 [结果/结论] 该方法能有效提高检索系统的查全率、查准率,证明了本文提出的技术的可行性、有效性。

关键词: 查询扩展 关联数据 激活扩散模型 DBpedia WordNet
分类号: G203
DOI: 10.13266/j.issn.0252-3116.2015.04.018

1 引言

查询扩展是信息检索中的一个热门的研究领域,其是针对单一查询词不准确的现状,对用户输入的相关实体属性查询在描述基础上进行语义层面的同义或近义等方面的扩展,利用新词来扩展初始查询,并在二次查询中给予词语重新加权,提高用户对检索结果的认可程度[1]。当前查询扩展技术主要可以分为基于传统关键词的查询扩展以及语义查询扩展。其中,传统关键词查询扩展方法又可以分为基于全局分析方法、基于局部分析方法、基于用户日志方法、基于关联规则方法;而语义查询扩展又可以分为基于通用本体方法、基于领域本体方法和基于关联数据方法。详细分类见图 1。

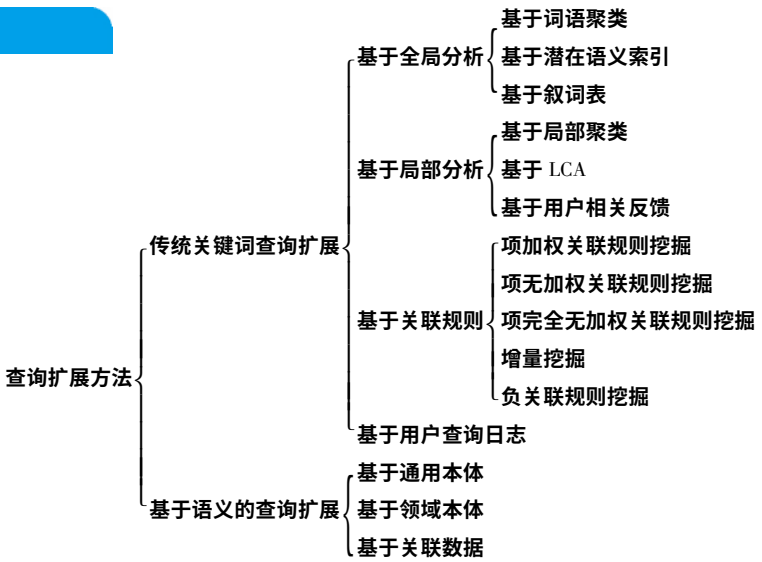

图 1 查询扩展技术方法分类

无论何种方法,扩展词的来源主要分为 3 种: ①利用共现分析等技术,从语料库中获得与初始查询词关联度较高的词语; ②利用聚类等技术,从部分语料库中获得与初始查询词关联度较高的词语; ③利用外部语义词典或本体如 WordNet、HowNet 等生成扩展词。

本文在关联数据融合的基础上,引入扩散激活模型(spreading activation model,SA) 理论,提出一种关联数据驱动的语义查询扩展方法,该算法在输入查询语句后,对提取的查询词的语义特征在知识库中进行有特定机制的扩散和激活,最后对这些语义关联的候补概念(candidate concept) 进行收集,并利用推理机制进行筛选,进而得到更优的概念集。利用 SA 模型,一方面保证了自动化的潜在语义扩展,另一方面有利于在扩展中利用语义信息进行推理。实验结果证明,该方法能够有效地提高信息检索的查全率和查准率。

作者简介: 田野(ORCID: 0000-0001-5335-2673) ,助理馆员,硕士,E-mail: ytian@lib.sjtu.edu.cn; 杨眉(ORCID: 0000-0002-4282-6738) ,副研究员,博士; 祝忠明(ORCID: 0000-0002-2365-3050) ,研究员,博士生导师; 张静蓓(ORCID: 0000-0002-2439-5049) ,助理馆员,硕士。
收稿日期: 2014-11-03 修回日期: 2015-01-15 本文起止页码: 122-128 本文责任编辑: 杜杏叶

2 关联数据驱动的语义查询扩展方法

2.1 相关研究及基本思想

利用关联数据进行查询扩展的研究目前还处于发展阶段,现有的研究主要是一些理论框架的阐述说明,鲜有完整的利用关联数据进行查询扩展的研究。如在文献[2]中,作者提出了基于DBpedia的语义扩展框架,整个系统分为两个步骤,首先是从知识库中提取候补概念,然后选出K-Best关联的概念,接着对被选中的K-Best概念,计算与查询词的相似度并排序,且该相似度算法是一种基于Wikipedia的语义分析进行的。文献[3]首先利用自然语言处理技术NLP,把查询语句转化为结构化的查询词,然后利用LOD(lirking open data)集中的图结构(graph structure)去关联有用的属性。

可以看出,传统的基于关联数据的扩展方法一般分为两种,两种方法都是首先把输入的搜索语句进行自然语言处理和预处理。具体来说,第一种方法[2]首先把输入的语句进行结构化和层次化处理,然后将这种结构化的形式带入到关联数据中去进行结构相似度的对比和扩展;另一种方法[3]是直接取出关键词,把关键词带入到知识库(关联数据)中,去关联出更多的潜在关键词,达到扩展的效果,最后对这些扩展的关键词进行排序以及选取排序靠前的关键词(Top-K)作为候选扩展词。无论采用哪种方法,都是利用自然语言处理的方式得到关键词的语义特征,然后利用这些特征或相似性算法等直接应用到LOD集中,其可扩展性受到很大的束缚,并且关键词容易扩展过度。

本文所提出的方法虽然也是一种基于关键词的潜在关联,但在带出潜在关键词的时候,不仅搜索出该实体的直接描述(第一层次的关联),并且通过带有衰减的扩散去激活更多层次的关联。此外,传统方法在候选词的选取上容易造成语义过度扩展,使得搜索变得陈杂,本文摒弃了传统的排序选词方式(Top-K),而是通过上下位推理法则的使用,一定程度上收敛了语义扩展过度现象。

综上所述,本文引入扩散激活模型(SA模型)理论,将该模型应用到关联数据图中,使得在输入查询词后,搜索潜在语义时,对提取的语义特征在知识库中进行有特定机制的扩散和激活。如此,一方面可以扩展出更多的潜在语义信息,另一方面也可以筛选出更加符合用户实际检索意图的关键词。

整体扩展方法步骤和流程如下:

- 步骤1:自然语言处理与去噪。
- 步骤2:初步选定搜索语句中的候补项。
- 步骤3:把选定的候补项带入SA模型,在关联数据中激活。
- 步骤4:为了避免扩展过度,把激活的多组潜在语义关键词进行统计和上下位推理。
- 步骤5:对筛选后的关键词再次进行语义搜索。

整个方法中核心部分为步骤3和步骤4,下文会具体介绍每个功能模块的职能。

2.2 查询扩展架构

2.2.1 扩散激活模型 扩散激活模型(SA模型)[4]起源于心理学,最初是研究人类记忆机制的模型,并被证明是一种具有高度解释力的模型。其搜索过程为:首先利用权值或激活值初始化一组节点集,然后通过反复迭代扩散到与所有源节点相链接的其他节点。在SA模型中,概念之间既有语义形式上的关联强度,也有概念本身的强度,当一个概念被刺激或被加工,该概念所在的节点便被激活,然后这种激活会沿着节点的各个连线向四周扩散,在扩散的过程中,首先扩散到与之直接相连的节点,再扩散到其余的节点。这种激活的数量是有限的,一个概念受到的加工时间越长,越有可能熟悉效应,并且该激活是遵循能量递进规律的。

后来SA模型被广泛应用于信息检索领域、人工智能甚至生物领域等,F. Crestani[5]指出了SA模型中的四点约束,包括距离约束、扇出约束、路径约束和激活约束。近些年来随着语义网技术的发展,该模型也被很多人应用到语义网络中进行语义扩展,其优势被很多学者认同[6]。文献[7]利用SA模型在本体知识库中扩展搜索同一领域的社团。文献[8]将该模型应用到语义网中,并根据概念的激活值进行语义扩展搜索。国内学者潘建国[9]在用户模型的结构特征基础上,提出了一种单向扩散激活的用户模型进化方法,该方法可以限制扩散的方向以及控制强度的衰减来更新关联节点,进而实现用户模型的进化,使得该用户模型可以及时反映用户的兴趣变化。

纯扩散激活模型非常简单[10],这种纯SA模型(见图2)是由网络数据结构组成的,激活的过程是通过在某处开始节点上放置激活权重,然后沿着开始节点的链接,激活权重开始扩展,激活过程为迭代进行,直到终止条件满足才结束。

SA模型的处理流程是通过一系列类的迭代序列定义的,迭代过程直至激发了终止条件,这种模型一次传递分为:①预调节阶段;②扩散阶段;③后期调节阶段3个阶段,见图3。

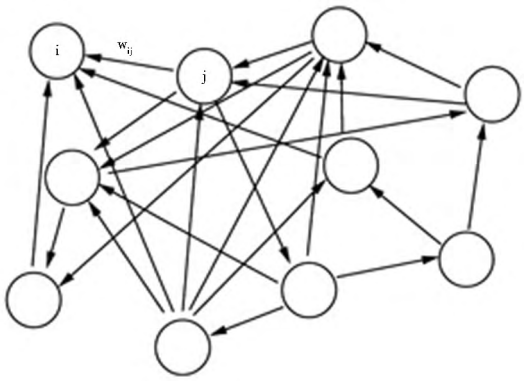

图 2 纯扩散激活模型

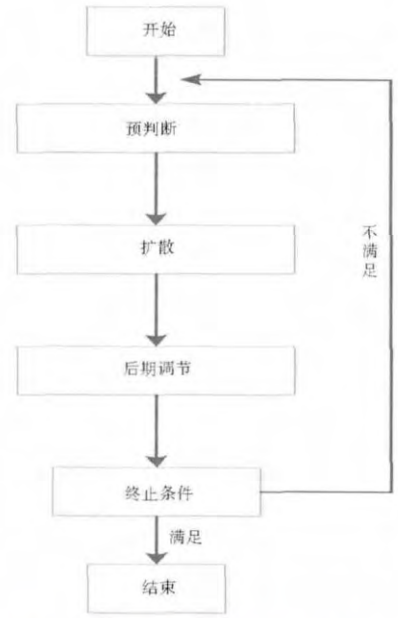

图 3 扩散激活模型的处理流程[10]

预调整与后期调整阶段都是可选的，这些阶段用以避免激活保留前次传递的状态，既控制单个节点的激活，又控制整个网络的激活。扩散阶段包含一系列从某个节点到所有连接的节点的激活传递。有许多扩散激活的方法，其中最简单的形式是在一个节点单元的情况下采用如下公式进行计算：

$$I_j = \sum_i O_i w_{ij} \quad (1)$$

其中，I_j 是节点 j 的总输入；O_i 是连接到节点 j 的单元 i 的输出；w_{ij} 是连接节点 i 和 j 的权值。

在节点计算完成其输入值之后，其输出值必须确定下来。通常情况下激活因子是关于该节点输入值的一个函数：

$$O_j = f(I_j) \quad (2)$$

当某个节点计算出其输出值后，它再继续向所有连接的节点进行扩散直到远离初始激活的节点，在一定数目节点的传递后需要进行终止条件检查，如果满足条件则扩散过程结束，否则继续下一次的传递。

本文提出的查询扩展方法就是将处理后的查询词带入该模型中，然后在关联数据图中进行激活，即利用 DBpedia 和 WordNet 对查询语句扩展提供语义支持。

2.2.2 查询扩展模型 本文在借鉴传统研究的基础上，引入 SA 理论，建立了一套激活机制和算法模型，对提取的关键词的语义特征放到关联数据的知识库中进行扩散和激活，在扩散和激活的过程中，搜集候补概念（candidate concept）时将产生很多概念类似但却完全不同的冗余概念，这里利用上下位的推理和语义相似度来选取有核心价值并且更优的概念实体。这种机制一方面确保了自动化的语义扩展，另一方面有利于充分利用其中隐含的语义信息。查询语句的语义扩展整体框架如图 4 所示：

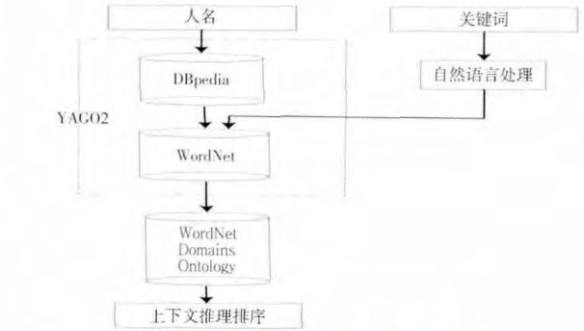

图 4 查询语句语义扩展框架

在该模型中，查询语句的扩散主要借助的是 DBpedia、WordNet、WordNet Domains 及 YAGO 知识库。

（1）WordNet。WordNet[11] 是美国 Princeton 大学开发的一个基于语言学规则的可计算的英语词汇知识库，其采用语义网的形式来表示词汇本体，即根据同义词集合（Synsets）来组织词汇，在其最新的 3.0 版本中，总共 155 287 个词汇被组成了 117 659 个同义词集。WordNet 中最主要的关系有同义关系、反义关系、上下位关系和部分/整体关系等，其中，同义关系是 WordNet 中最重要的关系。从 WordNet 的特性中可以看出，这种特性为查询词的语义扩充提供了很可靠的支持，但是在如此丰富的语料库中，抓取潜在语义就好比在大海中航行一样，如果不想失去方向，就需要一种良好的扩散机制来驾驭这种语义，这就是本文采用扩散激活模型的原因。

（2）WordNet Domains。WordNet Domains[12] 是一个在 WordNet 基础上开发的知识库。其试图将 WordNet 中有限的领域标签扩大到尽可能多的同义词集，即针对 WordNet 词典中的每个同义词的词集，利用人工进行标注，使得每个同义词集至少包含 1 个领域标签。

这些可用的领域标签大约有 200 个,以层次结构的形式进行组织。WordNet Domains Ontology 是基于 DDC 分类法,利用 Protégé 手工建立的本体。

(3) DBpedia。DBpedia 提供了百科全书式的科普描述功能,在输入某些研究人员的名字后,可以根据 DBpedia 进行匹配和扩展,推测和锁定用户感兴趣的话题。假设在 DBpedia 中搜索"Einstein"后,系统会自动扩展到其代表性的领域,比如相对论,然后"Phisics"等领域词汇就会被被语义关联上,因此可以快速锁定查询词的上下文背景(context)和与其相关的领域信息。

(4) YAGO。YAGO[13] 是一种融合了 DBpedia 和 WordNet 的知识库,并且将 WordNet 的高质量和 DBpedia 的高覆盖率相结合。相对于 DBpedia 大量的文章内容,YAGO 主要包含的是每个文章中的信息框和类别层次页面。YAGO 同样也包含实体(Entity)和关系(Relations),截至目前已经容纳了 170 万个实体和 1 500万个事实。在其最新的版本 YAGO2 中,针对时间和空间维度,增加了对实体、事实和事件等的标注;而在知识库资源方面,在增加了 GeoNames 后,YAGO2 一共包含了大约 8 千万个事实和 980 万个实体。

本文扩撒激活过程主要包括两个部分:一个是查询关键词的扩散;一个是人名的扩散。对于关键词这部分,可以直接在 WordNet 中进行扩散,进而得到一个个同义词集;对于人名这部分,主要采用的是设定起始点为 DBpedia,根据关联数据的特性,DBpedia 将会被自动链接到 WordNet 中的属性,继而扩散到具体 WordNet 中所对应的领域类别,最后甄别出该科学家的研究领域等背景信息。

2.2.3 查询扩展模型的扩散激活步骤 扩散激活算法的总体思路是:利用关联数据的属性(Predicate)来控制扩散的总体方向,再加上自身的扩散衰减,共同驾驭扩散的广度;在激活时,如果搜索到目标实体,则进行激活。因此该模型一方面利用了衰减因素来控制它的有限扩散,另一方面利用了关联数据的属性,即三元组当中的"predicate"性质来人为地控制它的扩散方向,这样可以显著提高结果集的有效性。

对关联数据图的扩撒激活步骤可以总结如下:

(1) 步骤 1:关键词的语义扩散。输入的关键词首先需要经过一系列的自然语言处理和预处理,这与前面消歧部分所提到的方法有诸多吻合之处,包括分词、词性标注、去停词、选词和去噪等过程,经过处理后所剩下的关键词直接放入 WordNet 中进行扩散,进而得到一个个同义词集合。

(2) 步骤 2:人名的提取、匹配和扩散。对于人名这部分,主要采用的是设定起始点为 DBpedia,根据关联数据的特性,DBpedia 将会被自动链接到 WordNet 中对应的属性,继而扩散到具体的 WordNet Domains 本体中。因此可以激活人物的相关属性,甄别其研究领域等背景信息。

相关的激发扩散方向可以设定为从"RDF: type"到"RDFS: subClassOf "最后到"YAGO: hasWordNetDomain",因此该规则(Rules)可以表示为:

——194206723 - Rule1:

(? entity RDF: type YAGO: wikicategory) →(? entity: direction RDF: type)

——194206722 - Rule2:

(? entity RDF: type YAGO: wikicategory) →(? entity : direction RDFS: subClassOf)

——194206721 - Rule3:

(? entity RDF: type YAGO: WordNet) →(? entity: direction YAGO: hasWordNetDomain)

接下来针对上述两个扩散过程进行收敛,即实体及其描述的激活:

- 放置激活权重(本文设定为 1.0,衰减值为 0.5);
- 利用权值初步化两组激活实体;
- 反复迭代直到实体类型是 WordNet Domains: a. 判断激活的实体类型和方向,b. 对于相关的实体,如果满足一组约束规则,则将其添加到激活集合中;
- 输出结果。

(3) 步骤 3:融合前两者,推理和筛选出高效的语义扩展集。融合上述两个过程中激活出来的候选扩展词,进行整理后,再经过简单的上下文推理,筛选出无重复的高效的语义扩展集。

可以看出,这种筛选后的携带语义的关键词,就将设定在特定的上下文背景中,然后在这些潜在语义信息的约束下,检索将变得更加精确,自然也就返回更多有意义的信息。

2.2.4 查询扩展模式 利用提出的语义扩展框架,在输入语句"data mining Jiawei Han"后,在知识库中(WordNet 和 DBpedia)进行语义扩散激活的过程见表 1。

3 实验与验证

实验环境是在 Win7 操作系统基础上,以 Java 作为开发语言、Eclipse 7.0 作为开发平台,利用 JDK 1.6.0.013版本,并在该环境下集成使用了 Jena2.7.4 及 Fuseki 工具等。

表 1 “data mining Jiawei Han”在知识库中语义扩散激活过程

候选项	潜在语义	筛选	语义扩展后的结果
Data	Data,datum,information,	Data,information	Data,information,
Mining	Mining,excavation,minelaying,mine	Mining,Extraction,minelaying	mining,extraction,minelaying,
Jiawei Han	Computer scientist	Computer science	computer science

3.1 数据选择

基于对据关联数据源描述信息的相关性及数据的可获得性等方面的考虑,本实验选择如下数据集(见表 2):DBpedia、WordNet、ACM、IEEE、DBLP、Citeseer,并对它们进行了关联数据的语义融合。

表 2 LOD 中融合的学术资源集

数据集	DBpedia	WordNet	ACM	IEEE	DBLP	Citeseer
数据量	>10 minion entities	117 thousand synsets	513 293 triples	38 700 triples	31 097 257 triples	3 085 153 triples

3.2 关联数据驱动的查询扩展实证

本节实证首先从系统的宏观层面说明查询扩展的效果,再从系统内部工作机制入手,详细阐述其工作的流程和效果。总体的系统界面如图 5 所示:

图 5 系统开发界面

当输入“data mining Jiawei Han”后,相关结果见图 6。

从图 6 可以很明显地看出,结果集中不仅包含了“data mining”,并且与之语义最相关的“association rules”“knowledge discovery”等也被抽取出来,因此从一定程度上证明了这种关联数据驱动的查询扩展方法的可行性和有效性。

下面详细阐述该系统的内部运行机制。

3.2.1 关键词的查询扩展检索 上述例子是从作者科研成果的全面描述的角度来进行的。本节将从关键词

图 6 对“data mining Jiawei Han”的查询扩展结果

的查询扩展角度来进行对比。本部分实验的基本思想是:利用关键词在单一本地数据集的检索结果与利用本文提出的查询扩展方法后的检索结果进行对比分析。

当从单一本地 Citeseer 数据集中检索“social semantic web”时,其查询语句和检索结果见图 7、图 8:

图 7 本地 Citeseer 数据集中关键词的 SPARQL 查询语句

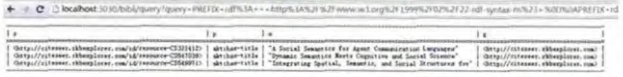

图 8 本地 Citeseer 数据集中关键词的检索结果

再次,修改查询语句,从融合后的数据集中利用本文提出的关联数据驱动的查询扩展方法进行检索,其 SPARQL 查询语句和检索结果展示见图 9、图 10。

从图 10 可以看出,在单一本地 Citeseer 数据集中检索“social semantic web”时,系统返回了 3 条结果集;而在利用本文提出的查询扩展方法后,首先是返回了来自 3 个数据源的检索结果,包括本地数据集 Citeseer 以及其他两个关联数据集 ACM 和 DBLP。此外,从前后对比的结果集中也可以看出,本文的关联数据驱动

图 9　LOD 集融合条件下关键词的 SPARQL 查询语句

图 10　LOD 集融合条件下关键词的查询扩展检索结果

的查询扩展从某种程度上排除了在本地数据集检索时造成的关键词“social semantic web”查询分散现象,这种 LOD 融合条件下的查询扩展结果集都是与输入关键词主题最相关的,包括“semantic web”和“social network”等。因此可以说,本文提出的关联数据驱动的查询扩展方法是具有一定的可行性和有效性的。

3.2.2　查询扩展实验结果分析　接下来的实验是对使用了本文提出的关联数据驱动的查询扩展方法前后的情况进行测试对比。测试的指标包括查全率、查准率和 F_1 值。用这些性能评价指标来证明本文提出的查询扩展方法的查询质量。测试结果如图 11 所示:

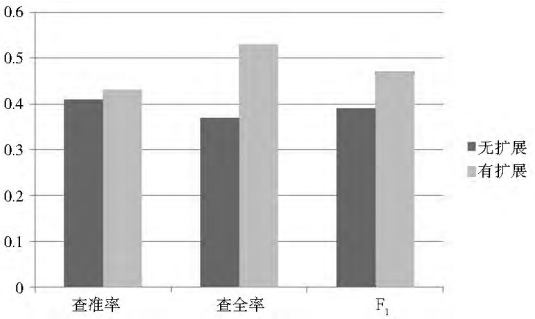

图 11　使用关联数据驱动的查询扩展方法前后查询性能的对比

实验结果表明,利用了本文提出的查询扩展方法后,系统的查全率和查准率都有所提高,整体 F_1 值也有相应的改善。正是由于对查询词进行了扩展,因此返回的结果集的数量必然增多,相应的查全率的提升是比较显著的。为了证明返回结果集的数量对整体系统查询性能(F_1 值) 的影响,下面将测试候选扩展词的数量对查全率、查准率的影响。

本节选取了前 30 个返回结果集作为验证数据源。在候选扩展词数量不同的情况下,查询扩展系统的查全率、查准率如图 12 所示:

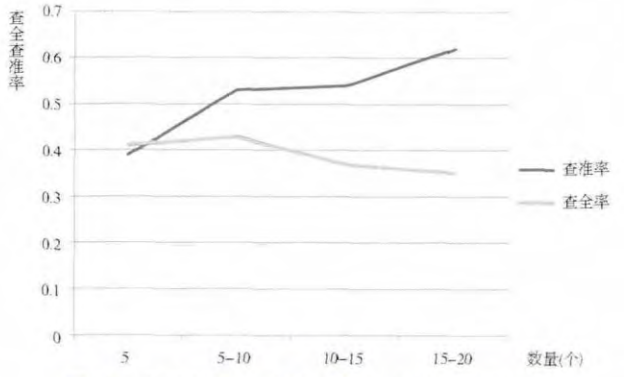

图 12　候选扩展词数量对查全率、查准率的影响

从实验结果中可以看出,随着候选扩展词数量的增多,查全率和查准率也有相应的波动。一般来说,扩展词的数量在 5 - 10 个之间时,其查全查准率,即查询性能达到最大值; 此后,随着扩展词数量继续增多,尽管查全率一直在变大,但是会加入不必要的噪声导致查询歧义,造成“查询偏移”现象,反而查准率整体下滑。因此,从一定程度上来说,候选扩展词的数量一般控制在 10 个之内时,其查询性能是最佳的。

此外,笔者除主观地判断查询的有效性外,还采用了简单的系统统计法。一般来说,查询扩展的结果,既包含了原有的直接与关键词匹配的结果,也有一些潜在的虽然没有直接匹配,但却是语义相关的结果。笔者邀请了 10 位计算机专业的研究生使用本系统,并对他们使用该系统进行假设,即如果在系统使用中点击了除输入关键词外的潜在推荐结果,就视为该扩展结果有意义。对这些人的总点击次数进行相关统计后,结果如图 13 所示:

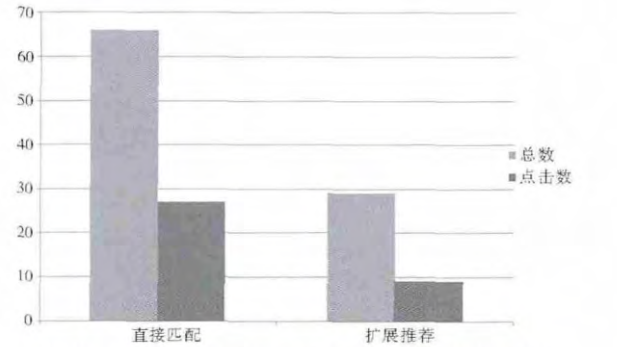

图 13　系统统计法的查询扩展结果有效性统计

从图 13 中可以看出,与输入关键词直接匹配的结果有 40% 被点击,而在扩展后的关键词的检索结果

中,同样有31%被点击,这说明潜在推荐的扩展词的相关度还是比较高的。

4 结论

在信息检索中,查询扩展是改善和提高检索系统性能的关键技术之一。本文提出了一种关联数据驱动的查询扩展方法,这种方法通过引入扩散激活模型理论,对提取的查询词的语义特征在知识库中进行有特定机制的扩散和激活,一方面利用了衰减因素来控制它的有限扩散,另一方面利用了关联数据的属性人为地控制它的扩散方向,显著提高了结果集的有效性。最后再对这些语义关联的候补概念(candidate concept)进行收集,并利用推理机制进行筛选,进而得到更优的概念集。利用SA模型,一方面保证了自动化的潜在语义扩展,另一方面有利于在扩展中利用语义信息进行推理。实验结果证明,该方法能够有效地提高信息检索的查全率和查准率,具有很实际的推广价值。

参考文献:

[1] Robertson S E, Jones K S. Relevance weighting of search terms [J]. Journal of the American Society for Information science, 1976, 27(3): 129-146.

[2] Aggarwal N, Buitelaar P. Query expansion using Wikipedia and DBpedia [C]//CLEF 2012 Evaluation Labs and Workshop. Rome: springer, 2012: 174-183.

[3] Augenstein I, Gentile A L, Norton B, et al. Mapping keywords to linked data resources for automatic query expansion [M]. Springer: Berlin Heidelberg, 2013: 101-112.

[4] 王甦,王安胜. 认知心理学 [M]. 1992. 北京: 北京大学出版社, 1992年.

[5] Crestani F. Application of spreading activation techniques in information retrieval [J]. Artificial Intelligence Review, 1997, 11(6): 453-482.

[6] Ziegler C N, Lausen G. Spreading activation models for trust propagation [C]//Proceedings of the 2004 IEEE International Conference. Taipei: IEEE, 2004: 83-97.

[7] Alani H, O'Hara K, Shadbolt N. Ontocopi: Methods and tools for identifying communities of practice [C]//Proceedings of the IFIP 17thWorld Computer Congress - TC12 Stream on Intelligent Information Processing. Montréal: springer, 2002: 225-236.

[8] Rocha C, Schwabe D, Aragao M P. A hybrid approach for searching in the semantic Web [C]//Proceedings of the 13th International Conference on World Wide Web. Rio de Janeiro: ACM, 2004: 374-383.

[9] 潘建国. 基于语义的用户建模技术与应用研究 [D]. 上海: 上海大学, 2009.

[10] Fernandez-Amoros D, Gil R H, Somolinos J A C, et al. Automatic word sense disambiguation using cooccurrence and hierarchical information [M]. Berlin: Springer, 2010: 60-67.

[11] Miller G A. WordNet: A lexical database for English [J]. Communications of the ACM, 1995, 38(11): 39-41.

[12] WordNet domains [EB/OL]. [2014-12-20]. http://wndomains.itc.it/.

[13] Nickel M, Tresp V, Kriegel H P. Factorizing YAGO: Scalable machine learning for linked data [C]//Proceedings of the 21st International conference on World Wide Web. Lyon: ACM, 2012: 271-280.

作者贡献说明:

田野:负责论文主体部分的撰写及系统开发;

杨眉:负责研究现状部分及系统后期开发;

祝忠明:提供论文思路及整体技术路线;

张静蓓:系统部分开发。

Research of Linked Data-driven Query Expansion

Tian Ye[1] Yang Mei[1] Zhu Zhongming[2] Zhang Jingbei[3]

[1]Shanghai Jiaotong University Library, Shanghai 200240

[2]The Lanzhou Branch of National Science Library, Chinese Academy of Sciences, Lanzhou 730000

[3]Shanghai International Studies University Library, Shanghai 201620

Abstract: [Purpose/significance] The current query expansion faced technology bottleneck, this paper presented a linked data-driven query expansion to improve retrieval system's recall precision. [Method/process] Applied the spreading activation model to the linked data graph. When input query words and searched for potential semantic meaning of query terms, there was a specific feature extraction mechanism of diffusion and activation in the knowledge base. Finally the candidate concepts for these semantic association were collected. [Result/conclusion] This method can improve retrieval system's recall precision. The technical feasibility and effectiveness was demonstrated.

Keywords: query expansion linked data spreading activation model DBpedia WordNet