软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2011,22(8):1771−1784 [doi: 10.3724/SP.J.1001.2011.03852]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

给互联网建立索引:基于词关系网络的智能查询推荐*

李亚楠[1,2], 王 斌[1+], 李锦涛[1], 李 鹏[1,2]

[1](中国科学院 计算技术研究所,北京 100190)

[2](中国科学院 研究生院,北京 100049)

Indexing the World Wide Web: Intelligent Query Suggestion Based on Term Relation Network

LI Ya-Nan[1,2], WANG Bin[1+], LI Jin-Tao[1], LI Peng[1,2]

[1](Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100190, China)

[2](Graduate University, The Chinese Academy of Sciences, Beijing 100049, China)

+ Corresponding author: E-mail: wangbin@ict.ac.cn

Li YN, Wang B, Li JT, Li P. Indexing the World Wide Web: Intelligent query suggestion based on term relation network. *Journal of Software*, 2011,22(8):1771−1784. http://www.jos.org.cn/1000-9825/3852.htm

Abstract: Search engine queries are often too vague to achieve relevant results. This paper presents an intelligent query approach that can distinguish vague queries and organize the related queries of each vague query into a concept hierarchy. Through the concept hierarchy, users can quickly find proper queries for their informational needs. The TECH (term concept hunting) is proposed, based on the small world of human languages. TECH utilizes both the community detection algorithms in the physical field and IR techniques in the computer science field to generate an extensible framework. Experimental results show that compared with the traditional listing query suggestion manner, users prefer the intelligent query suggestion. TECH can effectively distinguish vague queries and significantly outperforms the other three state-of-the-art hierarchical building systems statistically.

Key words: information retrieval; query suggestion; small world; community detecting; experiment design

摘 要: 搜索引擎用户经常提交意图模糊的查询,从而导致搜索失败.为此,提出一种检索交互方式——智能查询推荐,它可以自动辨别查询是否语义明确,并对模糊查询建立体现其不同语义概念的分类目录,这个目录将帮助用户快速定位到合适查询.为了实现智能查询推荐,提出了一种基于自然语言小世界性质的查询语义识别算法——TECH(term concept hunting).TECH 综合利用了物理学领域社区发现知识和计算机领域信息检索技术,给出了一种可扩展的算法框架.实验结果表明,与传统查询推荐方式相比,用户更喜欢智能查询推荐;TECH 能够有效地辨识模糊查询的不同语义概念,并统计显著优于 3 个知名的对比系统.

关键词: 信息检索;查询推荐;小世界网络;社区发现;实验设计

中图法分类号: TP311 **文献标识码**: A

如今,我们通过向搜索引擎查询从互联网获取所需信息,但这种方式并不符合人类的习惯.我们可以将整个

* 基金项目: 国家自然科学基金(60603094, 60776797); 国家重点基础研究发展计划(973)(2007CB311103); 国家高技术研究发展计划(863)(2006AA010105); 北京市自然科学基金(4082030)

收稿时间: 2009-06-10; 定稿时间: 2010-03-08

© 中国科学院软件研究所 http://www.jos.org.cn

互联网上的信息看作一本没有目录的书,每一页的内容以一个具体查询作为“标题”.查找互联网信息检索就像是读这本书,我们可以直接翻到书的任何一页,就像通过搜索引擎我们可以直接看到任意一个查询的匹配结果.我们一页一页地浏览却不知道这些页面间的联系,不知道哪些“标题”属于同一个主题,不知道它们组成的篇章结构是什么.每次我们翻到一页(构造了一个查询),查看它是否包含我们需要的内容.如果我们没有翻到正确的页面,那么把书合上,根据我们的知识和经验再翻一次(查询重构),看这次的页面是否正确.然而研究[1]显示,用户查询经常含有歧义或意图不清,这导致用户经常搜索失败.

在现实世界中我们查找信息时,目录索引是不可或缺的重要信息,这些索引给我们带来了很多方便.我们能否给互联网之书也加上目录来帮助我们翻阅?直接对整个互联网信息构建统一的语义目录是非常复杂和难以实现的,而且也是没有必要的.实际上,每个用户只关心与其查询意图相关的那部分目录,即相关查询之间的联系及它们所属的语义概念.基于上述思想,本文给出一种信息检索交互方式——智能查询推荐.其对不同类型的查询给予不同的推荐方式:如果查询含义明确,那么直接列出与其相关的其他查询词;否则,智能查询推荐将根据用户查询的各种不同相关语义概念将推荐词整理到不同的概念目录下.通过智能查询推荐,用户不用费力去思考查询词,他只需要浏览层次化的推荐目录并选择更合适的查询.

智能查询推荐不同于当前被广泛使用和研究的查询推荐[2,3],它们都是直接罗列几个相关查询.由于查询推荐的目的在于猜测用户的各种可能的意图并给予帮助,因此召回率对于查询推荐更为重要.直接罗列的方式难以同时满足高召回率和低浏览负担.与查询扩展[4,5]相比,智能查询推荐不需要扩展原始查询而增加系统开销[6];而且研究表明,用户更喜欢交互式的查询推荐而不是自动进行的查询扩展[7].很早就有人研究在信息检索交互中引入层次化目录,并证实这样确实可以增强用户体验[8].这些工作包括早期的基于文档集聚类的方法[9]、近几年的查询结果聚类[10,11]及 Hierarchical Faceted Search Interfaces[12].这些方法主要基于文档聚类,它们假设用户查询正确并帮助用户快速浏览查询返回文档.而查询推荐用于在用户查询不准确时帮助用户重构查询.智能查询推荐需要直接对查询词进行处理并挖掘出查询词间的语义关系,而不是检索结果文档间的层次关系.

相对于文档,查询非常短且可利用的文本特征有限,传统的检索算法往往不适用于查询.与基于文档聚类的层次化目录生成问题相比,识别查询词间的语义关系并实现智能查询推荐似乎更具挑战性.智能查询推荐关键需要解决:(1) 辨别一个查询是否意图模糊;(2) 识别模糊查询的不同语义概念并分类标识各个相关查询词.本文提出了一种基于词关系网络小世界性质的查询词语义识别算法,我们将其称为 TECH(term concept hunting). TECH 可以同时解决上述两个问题.这里,词关系网络是由词或短语及它们之间关系构成的图,节点表示词或短语,存在一定联系的词或短语用边相连.研究[13,14]显示,人类语言的词关系网络存在小世界效应.我们利用了这一自然规律,并提出一种有效的技术解决方案.TECH 是一个可扩展的算法框架.它综合利用了复杂网络领域的社团发现算法和信息检索技术,并给出了效果显著的推荐结果.

基于 TECH,我们实现了一个智能查询推荐原型系统——Jigsoo.智能查询推荐的评价也是一个开放性的问题,我们定义了一套评价准则并开展了用户交互评测.实验结果显示:相对于传统查询推荐——直接罗列相关查询,用户更愿意接受智能查询推荐;与两个知名的商业搜索引擎层次化目录相比,用户更喜欢 Jigsoo 推荐的结果;对比 SIGIR 2007 中提出的一个层次化目录生成算法[11],我们的算法能够提供更好的结果.Jigsoo 只是基于 TECH 的一种简单实现,通过将更多有效的资源或算法融入 TECH 的算法框架,我们相信智能查询推荐将取得更加振奋人心的结果.

1 智能查询推荐

查询推荐是一种能够有效提高用户搜索体验的信息检索交互技术[7],而且现在已经广泛被各大搜索引擎采用.近几年,查询推荐已经成为信息检索领域的一个热门方向[2,3].这些研究工作仍旧采用传统的查询推荐方式——直接将相关查询简单罗列,然而这一方式存在两个缺点:

- 查询推荐的目的在于推测用户各种可能的意图,推荐查询应该尽量涵盖各种可能的查询,因此召回率就显得更为重要.为了不加重用户浏览负担,传统的查询推荐方式只能在推荐列表中放 10 个左右的查

© 中国科学院软件研究所 http://www.jos.org.cn

询,但是对于那些含义宽泛的查询,这样难以获得较好的召回率.

- 另外,当查询模糊时,推荐结果往往都只与其中最流行的那个概念相关,而想表达其他含义的用户得不到任何帮助.例如,用户想了解关于火箭的科技或政治内容(例如,朝鲜火箭发射),输入查询“火箭”,然而几乎所有搜索引擎返回的结果都是与 NBA 篮球有关的推荐查询.

智能查询推荐克服了这些缺点并提供了一种良好的交互方式:通过智能地区分清晰查询和模糊查询并对模糊查询建立体现不同语义概念的分类目录,既保证了高召回率又符合用户的浏览习惯而又不加重检索负担.图 1 给出了智能查询推荐原型系统 Jigsoo 对歧义查询“sun”的推荐示例,“sun”可以表示太阳系,也可以指一个计算机公司或其他概念,智能查询推荐会自动识别这些不同的语义概念,并将推荐查询分类组织到相应目录中(图 1 中“computer”类已展开).如果用户输入的是一个含义明确的查询:“Sun Microsystems company”,那么智能查询推荐会发现该查询已经语义具体,直接列出与其相关的其他查询词:“java”,“Sun Microsystems company business”,“Oracle”,“IBM”等.

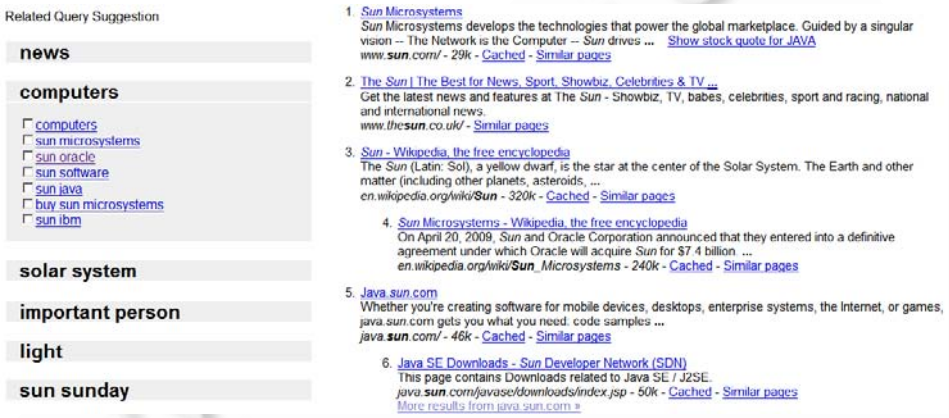


Fig.1 Intelligent term suggestion for vague query “sun”

图 1 智能查询推荐对模糊查询“sun”的推荐结果

智能查询推荐向用户展现了与其可能意图相关的各种相关查询,并根据用户查询模糊与否将其相关查询进行有效地组织.用户可以对各种相关查询间的语义联系一目了然,利用推荐查询从大量信息中获取所需的信息.如果将智能查询推荐看作是为海量的互联网信息建立了一种目录索引,在一定意义上讲,这种方式甚至好于传统的静态目录——涵盖所有信息的统一目录(如每本书开头关于全书内容的目录).首先,通过自动辨识模糊查询和构建语义概念目录,智能查询推荐极大地降低了用户浏览负担,用户不需要首先在总目录中定位相关目录.其次,不同于静态地按照编辑认知方式构造的静态目录,每次智能查询推荐都以当前查询为中心构建目录,这样更适合用户进行浏览.

尽管智能查询推荐拥有诸多好处,但如何将其实现却是个极具挑战性的问题.相对于文档对象,查询包含的文本内容很有限(平均长度为 3 个词左右[15,16]),不利于使用现有的信息检索、文本聚类和分类算法对其进行处理.而智能查询推荐却要求对查询词作语义层面上的处理,这里主要有 3 个关键问题需要解决:

i. 如何判断一个查询是否含义明确?这决定对该查询的推荐方式.

ii. 如果用户查询意图模糊,那么与该查询相关的语义概念有几个?

iii. 对体现不同语义概念的查询含义或话题,如何找出其相关查询?

当然,要实现一个完整的智能查询推荐系统,还需要解决相关查询的抽取和排序、层次化目录的构建等问题.本文主要关注于如何解决上述 3 个关键问题,其他问题的相关算法在现有的查询推荐、查询扩展、查询结果聚类等领域的大量文章中已经被大量阐述,读者可以方便地获取到有效的算法.

© 中国科学院软件研究所 http://www.jos.org.cn

2 查询词语义识别算法 TECH

为了实现智能查询推荐,我们提出了一种基于词关系网络小世界性质的查询词语义识别算法,称为 TECH (term concept hunting).TECH 算法可以同时解决上述 3 个问题.本节将详细阐述 TECH 算法.

2.1 直观意义

晋朝文学家傅玄有句名言:“同声自相应,同心自相知”,兴趣相同的人会相互结识.同样地,人类语言中语义相关的词也会经常在一起出现.西方也有句谚语:“要了解一个人,只要看看他所交的朋友”,类似地,一个词与其他词之间的关系也反映了这个词的语义概念.相对于文档,查询自身包含的文本特征很少,所以应用传统的基于文本特征的信息检索或分类聚类算法存在诸多困难.但是,我们却可以从众多包含文本的数据中找出查询间的关系,例如,词在文章中的共现关系、查询在搜索日志中的重构替换关系等.由词及它们之间的关系,我们可以构建一个词关系网络,其中,节点表示词,而边表示词之间的关系.物理学方面的研究显示:和朋友关系网络一样,人类语言的词关系网络也满足类似性质(小世界效应)[13,14,17].基于这一自然规律,我们可以利用一个查询的相关词关系网络分析这个查询的语义概念,就像我们可以利用一个人同其周围人的关系了解这个人.

小世界网络模型具有高聚团特性和小的平均最短距离[18],对应于现实的朋友关系网络,即广泛存在的各种朋友圈子和著名的六度分隔理论**.更直观地讲,朋友关系网络存在这样的特性:一个人的朋友很可能也相互熟识,这些人都在一个朋友圈里或居住在同一个地方,他们组成的关系网络形成了一个连边密集的社团(community),类比于词关系网络,就是同类词语会构成的一个由社团表征的语义概念;但是,也有些人会结识一些在其他朋友圈或异国他乡的人,这些人的存在解释了六度分隔理论,他们的朋友会分布在不同社团中,类比于词关系网络,就是模糊查询的相关查询词会构成多个社团.

词关系网络也存在社团结构[19],TECH 通过探测社团结构来辨别查询是否模糊,以及识别模糊查询的不同语义概念.对于一个查询,如果其相关词属于多个语义概念,那么该查询的相关词关系网络具有明显的社团结构(即该网络由若干个社团组成,每个社团内有密集的连边,而社团之间的连接却比较稀疏),其不同的语义概念由不同的社团表示,如图 2 所示(与“Java”相关的 3 个语义概念中的词聚成 3 个社团).如果查询含义清晰,那么其网络不具有明显的社团结构.

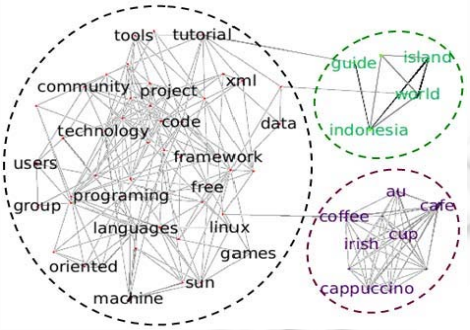


Fig.2 Term relation network about “java”

图 2 查询“Java”的词关系网络

直观上看,辨识一个查询是否语义模糊及识别其不同语义概念可以转化为一个图划分问题.然而,传统的图划分方法却并不适用于这一问题.传统的图划分算法基于这样一个目标[20]:寻找一种划分方法,使得不同类型节点间的连边数或连边加权和最小.尽管这类图划分方法在很多图上取得了很好的效果,但对于真实复杂网络的

** 20 世纪 60 年代,美国哈佛大学心理学家 Milgram 根据社会实验给出一个推断:地球上任意两个人之间的平均距离是 6.也就是说,平均中间只要通过 5 个人的传递,一个人就能与地球上任何一个角落的另一个人建立联系.

© 中国科学院软件研究所 http://www.jos.org.cn

社团发现却不是最适用的方法[21],因为真实复杂网络往往具有一些特殊性质,包括小世界性质.词关系网络是一种带有小世界性质的复杂网络[13,14].针对其这一性质,我们引入数学模型模块度(modularity).关于什么是词关系网络和模块度,我们将在下一节详细说明.

2.2 词关系网络及小世界性质

定义(词关系网络). 词关系网络 G 表示为二元组:$\langle T,A\rangle$,其中,$T=\{t_i\}_n$ 为网络中的词或短语集合,t_i 表示对应于节点 i 的一个词或短语(也可以是一个查询);$A=\{A_{ij}\}_{n\times n}$ 为词关系网络 G 中的邻接矩阵,如果 $A_{ij}=1$,则表示节点 i 与 j 邻接,即词 t_i 与 t_j 之间存在联系;如果 $A_{ij}=0$,则表明词 t_i 与 t_j 之间缺乏足够强的联系.这里的联系可以是辞典中的同义词关系、文档中词的共现关系等.

词关系网络存在社团结构[19],关于寻找网络社团结构的算法有很多,各种算法的具体内容参见文献[18,22].本文主要介绍和使用基于模块度(modularity)[23]的方法,这类算法具有直观的物理意义,准确度比较高,近几年已经成为社团结构分析的一种标准算法[22].模块度是一个用于度量网络社团结构显著性的质量函数,模块度的值越大,说明社团结构越明显.令 Q 表示一个无权无向图 G 模块度,给定一种社团划分,定义 e_{ij} 表示连接社团 i 内节点和社团 j 内节点间连边在全部边中所占的比例,则

$$Q=\sum_i (e_{ii}-a_i^2) \tag{1}$$

其中,$a_i=\sum_j e_{ij}$.

显然,e_{ii} 表示社团 i 内边在全部边中的比例,也就是真实状态下社团 i 与其自身连接的概率;a_i 表示所有社团与社团 i 中节点的连边在全部边中的比例,也就是社团 i 与其他所有社团的连接概率;进而,a_i^2 给出了在完全随机状态下拥有连接概率 a_i 的社团 i 与其自身连接的期望值.

模块度的物理意义是:如果我们把社团结构看作是由网络中存在不同类别的对象导致的,同类别节点间的连接概率要大于任意节点间的连接期望值,模块度就是网络中同类节点间联系所占的比例减去无类别随机网络在同样社团划分下的该比例期望值.很容易理解,模块度值越大,该网络与随机网络的差别就越大,其聚团性和分类性就越好,该网络的社团结构也就越明显.

大量研究[13,14]表明,词关系网络具有小世界性质.小世界网络具有两个重要性质:小的平均最短距离和高聚团性.具有小世界性质的复杂网络模型是小世界网络模型,由 Watts 和 Strogtz 在 1998 年的《Nature》上提出[17].小世界网络模型可以看作从完全规则网络向完全随机网络的一种过渡网络模型(简称 WS 模型).WS 模型可以这样构造[18]:考虑一个规则网络,网络中的节点围成一个环,其中每个节点都与它左右相邻的 $K/2$ 个节点相连.我们以概率 P 随机地重新连接网络中每个边,这样就构造出一个小世界网络.

因为小世界性质,基于模块度的算法比传统的图划分算法更适合词关系网络.对应词关系网络,最好的社团划分并非仅仅需要做到使社团间的连边最少,更重要的是要寻找社团间连边少于随机网络期望值的划分[21].关于这两类算法,更多理论和实验上的分析比较参见文献[21,24].

2.3 算 法

本节将介绍 TECH 及其中使用的社团发现算法.TECH 是一种可扩展的算法框架.它基于查询相关词关系网络,这种网络具有一些特殊的属性,原始的基于模块度的社区发现算法并不适合.

我们用有权有向图 $G_q=\langle R,W\rangle$ 表示查询 q 的相关词关系网络.其中,$R=\{t_1,t_2,\ldots,t_n\}$ 表示 q 的相关词或相关查询集合,t_i 为一个查询词或查询;$W=\{W_{ij}:1\leqslant i\leqslant n,1\leqslant j\leqslant n\}$ 表示边的权重,权重 W_{ij} 体现了从 t_i 到 t_j 之间的关系强弱,$W_{ij}=0$ 表示边 (i,j) 不存在.不同于一般的词关系网络,G_q 是一个有权有向图,这样可以更精确地刻画查询词之间的联系.例如,考虑查询“海淀医院”和查询“北京 医院”,搜索“海淀医院”的人很可能也想知道北京还有哪些医院可以选择,但是查询“北京 医院”的人却只有很小的概率关心“海淀医院”,因为北京有很多医院,这两个查询间的关系强弱是与其方向有关的.

© 中国科学院软件研究所 http://www.jos.org.cn

2.3.1 算法框架

TECH 包括 4 个步骤:获取查询相关词、构建查询相关词关系网络、探测网络的社团结构并划分、根据不同情况给予推荐.图 3 给出一个示例说明,这是一个实现智能查询推荐的技术框架,而不是具体的实现内容,因为在各步骤中,都有很多适合的算法可以选择.下面分别介绍算法框架包括的各个步骤:

i. 获取与当前用户查询相关的查询词或查询,即生成集合 R.在查询推荐、查询扩展等领域有很多方法可用于生成相关查询或查询词.例如,使用伪反馈的方法从检索返回结果中抽取查询相关词[5],从搜索引擎查询日志中挖掘相似查询[6].

ii. 构建由查询相关词组成的词关系网络,即确定集合 W.W_{ij} 可以看作是求解 t_i 和 t_j 间的相似度 $sim(t_i,t_j)$,可以针对不同任务通过不同的相似度函数 $sim(t_i,t_j)$.例如,利用互信息衡量文章中相邻词之间的关系,利用关联规则挖掘度量查询日志中查询间的关系.

iii. 利用模块度判定词关系网络是否有明显社团结构,有则进行社团划分.这一步需要分析词关系网络 G_q 并找到其最大模块度值 $Q_{\max}$.这一步是算法的关键,我们需要改进模块度函数,使其满足 G_q 以及智能查询推荐的一些要求,第 2.3.2 节将对其进行详细阐述.

iv. 对相关查询或查询词进行排序并分情况推荐.如果用户查询 q 含义明确,则直接对所有相关查询或词进行相关度排序,推荐排名靠前的结果;如果查询 q 意图模糊,则为每个语义概念社团打一个类别标签,并对社团内的相关查询或查询词进行排序,只保留排名靠前的结果.最后,各个社团也要进行排序.这一步的主要问题有两个:生成类别标签和排序.关于这些问题的具体解决办法,在层次化文档聚类[9,10,12]中有很多可用的方法.

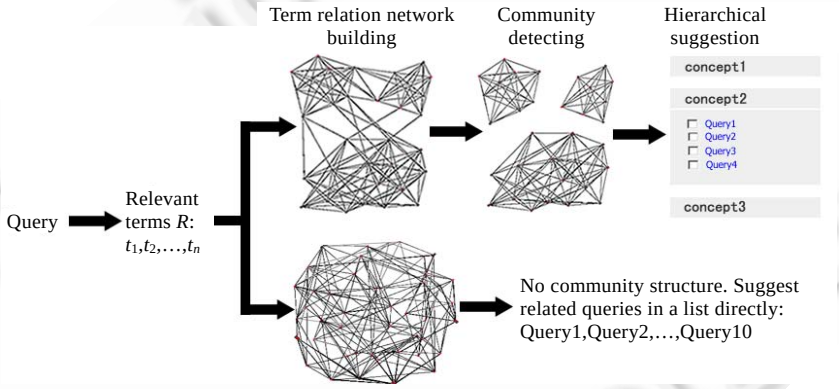


Fig.3 Framework of TECH

图 3 TECH 的算法框架

这是一个灵活的算法框架,具体实现方法可以有很多种.本文主要关注于其中关键的第 iii 步,第 2.3.2 节将给出一种有效的社团发现算法.

2.3.2 查询相关词关系网络的语义社团划分算法

原始的模块度函数(公式(1))用于无向无权图,并不适用于词关系网络 G_q.当然,可以将 G_q 转化成相应的无向无权图,然后再利用基于模块度的社团发现算法对其进行处理.但是,这样的社团划分存在两个主要弊端:

(1) 当两个关系很弱的社团间有很多边时,这两个社团会合并成一个社团,而两个拥有很强关系但连边数相对较少的社团则不会合并.

(2) 最终产生的社团大小具有幂律分布,即会生成一个很大的社团和很多很小的社团,这往往并不符合现实情况.

针对这些问题,我们改进了模块度函数,使其可以利用 G_q 中信息检索技术求出的词关系,同时避免上述两

© 中国科学院软件研究所 http://www.jos.org.cn

种缺陷.下面我们对其进行具体阐述.

首先,考虑一般的词关系网络 $G=\langle T,A\rangle$.回顾其定义(见第 2.2 节)我们知道,G 是一个无向无权图,A 为其邻接矩阵.令 k_v 表示节点 v 的度,m 表示 G 中无向边的数目,根据公式(1),我们可以将 G 的模块度函数写作

$$Q=\sum_{v=1}^{n}\sum_{w=1}^{n}\left[\frac{A_{vw}}{2m}-\frac{k_vk_w}{(2m)^2}\right]\delta(c_v,c_w),\ \text{即}Q=\frac{1}{2m}\sum_{v=1}^{n}\sum_{w=1}^{n}\left[A_{vw}-\frac{k_vk_w}{2m}\right]\delta(c_v,c_w) \tag{2}$$

其中,c_v 表示节点 v 所属的社团;函数 $\delta(c_v,c_w)$ 指示节点 v 和 w 是否在一个社区内.即当 $c_v=c_w$ 时,$\delta(c_v,c_w)=1$;否则,$\delta(c_v,c_w)=0$.

然后,考虑查询相关词关系网络 G_q.先对 G_q 所有的边权重进行归一化,令 $\overline{W_{vw}}$ 表示归一化后节点 v 和 w 之间的权重,则

$$\overline{W_{vw}}=\frac{W_{vw}}{\sum_a W_{va}} \tag{3}$$

利用归一化后的边权重修改模块度函数(2),可以得到:

$$Q=\frac{1}{n}\sum_{v=1}^{n}\sum_{w=1}^{n}\left[\overline{W_{vw}}-\frac{1}{n}\sum_{a=1}^{n}\overline{W_{aw}}\right]\delta(c_v,c_w) \tag{4}$$

函数(2)中表示边的 A_{vw} 被归一化后的加权值 $\overline{W_{vw}}$ 取代;无向图中的边数的 2 倍 $2m$ 被 n 代替,因为

$$\sum_{v=1}^{n}\sum_{w=1}^{n}\overline{W_{vw}}=n \tag{5}$$

两个节点度的乘积 k_vk_w 变成了 $\sum_a\overline{W_{aw}}$,这是由于

$$\sum_{b=1}^{n}\overline{W_{vb}}\sum_{a=1}^{n}\overline{W_{aw}}=\sum_{a=1}^{n}\overline{W_{aw}} \tag{6}$$

$\overline{W_{vw}}$ 可以看作一种条件概率 $P(t_w|t_v)$,即在词 t_v 出现的情况下 t_w 出现的概率;而 $\sum_a\overline{W_{aw}}/n$ 可以认为是词 t_w 出现的期望概率 $P(t_w)$.因此,如果两个词 t_v 和 t_w 经常在一起同时出现,而且出现概率高于随机期望值,则 G_q 的模块度 Q 会越来越高,从而 t_v 和 t_w 会更可能被聚到一个社团中.

与原有的模块度函数相比,等式(4)可以在一定程度上解决原有的两个弊端.对比等式(2)和等式(4),最大的变化在于方括号内.方括号内形成的差值结果越大,词 t_v 和 t_w 所在的社团 c_v 和 c_w 就越倾向于合并于同一个语义社团概念中,因为那样会取得更高的模块度值.在等式(4)中,被减数 A_{vw} 变为归一化的权重 $\overline{W_{vw}}$,这使得其间拥有高权重连边的词或社团会合并到一个大的社团中,而不是其间拥有很多低权重连边的社团.由于词关系网络中节点的度服从幂律分布[13,14],因此含有巨大度值的节点的社团往往会有大量的边与其他社团内的节点连接.根据等式(2),这些社团会吞并其周围的小社团,进而形成一个巨大的社团.边权重归一化后,度很大的节点的多数连边权重将很低,从而避免了上述问题.另一方面,如果 G_q 中所有的边的权重都为 1,则等式(2)方括号内的差值可以看作

$$A_{vw}-\frac{k_vk_w}{2m}=k_v\left[\frac{A_{vw}}{d_v}-\frac{k_w}{2m}\right]\approx k_v\left[\overline{W_{vw}}-\frac{1}{n}\sum_a\overline{W_{aw}}\right] \tag{7}$$

当 t_v 的相关词数量很多,即 k_v 很大时,等式(4)中的差值一般更小.这样就避免了更多的语义概念社团被合并进一个巨大的社团,从而使结果更符合实际情况.

寻找使模块度最大化的图划分方式是一个 NP 问题,这里可以采用一种贪心优化算法[25]求得近似最优解.开始时,假定每个节点属于不同的社团,该算法不断合并社团,直到找到一种使模块度值达到峰值的社团划分模式.每次合并时,选取能够使合并后模块度值增量最大的两个社团.该算法的时间复杂度为 $O(md\log n)$,其中,d 为合并过程树状图的高度.真实情况下,图往往比较稀疏,可以认为 $d\approx\log n, m\approx n\log n$,所以算法复杂度也可看作 $O(n\log^3 n)$.由于 G_q 最多只包含上百个节点,因此 TECH 可以很快地处理完一个查询.

© 中国科学院软件研究所 http://www.jos.org.cn

3 智能查询推荐原型系统 Jigsoo

根据 TECH 算法框架,我们搭建了一个简单的智能查询推荐原型系统——Jigsoo.下面我们简要介绍其实现过程.需要注意的是,Jigsoo 只是用于验证和评价 TECH 的简单原型系统.在 TECH 算法框架下,利用现有信息检索技术完全可以实现更为复杂、有效的智能查询推荐系统.

我们使用伪相关反馈的方法来获取每个查询的相关词.对每一个查询,我们获取 Google 的返回结果摘要进行预处理:去停用词、词根还原、对汉语分词、短语抽取,然后从中抽取出现频率最高的 100 个词或短语作为相关词.另外,我们还从 WordNet 中抽取了英文查询的相关词,并一同加入到查询相关词集合 R 中.我们根据查询相关词间的共现关系来建立词关系网络 G_q,两个词 t_v 和 t_w 在同一个结果摘要中出现的次数作为权重 W_{ij}.由于 G_q 中有很多噪音边,我们用 Jaccard 相似度(<0.015)和 Dependence 相似度(<0.15)过滤掉了相似度太低的边[2].对 G_q 中的社区划分采用公式(4)及贪心优化方法[25].根据文献[23],当模块度值在 0.3~0.7 之间时,表明社团结构明显,因此以 G_q 的模块度是否大于 0.3 来判断查询 q 是否模糊.如果 q 为模糊查询,G_q 将被划分成多个语义概念社团.对其中的每一个语义概念社团 C_i,我们利用了 ODP(http://dmoz.org)和 WordNet 中的树状目录生成 C_i 的类别标签:首先,我们尝试在 ODP 和 WordNet 中找到能够涵盖 C_i 中多数相关查询的目录名作标签.如果查找失败,我们就在 C_i 中找一个相关词或短语作为类别标签.作为标签的相关词需要与 C_i 中绝大多数词有联系且与 C_i 以外节点联系较少.

下面说明 Jigsoo 的处理效率.Jigsoo 使用 Python 实现,在一台拥有双核 3.0GHz Intel CPU 的 PC 上,我们使用 100 个查询对其测试.不考虑抓取 Google 结果的网络延迟时间,Jigsoo 平均处理一个查询的时间为 1.21s,其中预处理约 1s,后续 ODP 检索时间 0.2s,词关系网络社团划分时间的平均时间为 0.00 03s.在实际搜索引擎中,各种预处理工作是离线完成的,而且对 ODP 和 WordNet 的检索完全可以通过事先加入内存以及更好的索引加快速度.此外还应该注意到,Jigsoo 是用一种低效率的高级语言 Python 实现的原型系统.如果替换为 C 等高效率语言,在真实信息检索系统中的检索时间则完全可以应付查询实时处理.

4 实 验

本文实验的主要目标在于评价我们的方法对查询各种语义概念的识别效果,评测智能查询推荐的可用性(第 4.3.1 节)、TECH 算法辨识模糊查询(第 4.1 节)和不同语义概念(第 4.2 节)的效果.实验基于简单原型系统 Jigsoo,一个大规模的用户交互评测(user study)实验(第 4.3 节).关于推荐结果的准确率与召回率,Jigsoo 所用的伪相关反馈方法已被前人分析过.另外,根据 TECH 算法框架,相关词选取可以换用更好的方法.因此,尽管推荐查询的准确率和召回率也是很重要的评价指标,但却不是本文工作的评价重点.

对于智能查询推荐,没有标准的评价集.要做到客观而公正地判断一个查询是否模糊和识别其包含的各种语义意图是很困难的,即便人工评测,不同人的评价也经常不一致[1].为了保证评价结果的准确性,每个查询的结果需要多个用户从多个方面进行评价.这种实验代价较高,因此难以做到大规模评测.尽管测试查询不多,我们仍然取得了具有统计显著性的结果.

4.1 TECH能否辨识模糊查询

尽管很多研究都指出搜索引擎查询中存在大量的模糊查询[1,26],但是对于什么是模糊查询却没有统一的评价方法和标准.为了有一个准确的评测集,我们用文献[26]中给出的分类查询标注作为评价基准.在该文中,Nguyen 和 Kan 开展了一项用户调查研究.他们从 AllTheWeb 欧洲搜索引擎 2002 年 5 月 28 日的查询日志中随机抽取了 75 个查询,然后召集了 25 个志愿者对这些查询进行分类.其中,有一项分类任务是根据查询的模糊程度对查询分类:一个查询首先被判断是否含有歧义,如果有歧义则属于 polysemous 类;否则,该查询将根据其意图明确程度给予一个分值,分值范围为一个从 1(specific)到 5(general)的李克特量表(Likert scale).即,一个查询获得的得分越高,其含义就越是被认为笼统(general).实验中,每个查询受到 5 个志愿者的标注,最终得分为 5 个标注的平均分.文献[26]中给出了部分查询(28 个)的标注结果,我们选用其中得分为 1.0 的 6 个查询为含义明确查

© 中国科学院软件研究所 http://www.jos.org.cn

询标准集,选取其中的 polysemous 类 5 个查询和得分大于等于 4.0 的 8 个查询为模糊查询标准集(共 13 个查询).

我们以每个查询对应的词关系网络的模块度值(TECH 算法求得的最大值)作为判断一个查询模糊与否的标准,根据文献[23],阈值选取 0.3.表 1 给出了 Jigsoo 对模糊查询和明确查询的分类结果.实验结果显示,polysemous 类的 8 个查询全部被正确识别为模糊查询,general 类的 5 个查询中有 4 个查询被正确识别为模糊查询,specific 类的 6 个查询中有 4 个被正确识别为明确查询.从结果来看,Jigsoo 更倾向于避免把模糊查询错分为明确查询.对查询推荐而言,能够找到用户各种可能的意图更重要,因此用户更乐于接受将一个明确查询的相关查询按照分类目录方式组织,而不是将一个模糊查询误判为明确查询却漏掉一些相关查询推荐.判断失败的 3 个查询为:general 类的 presidents、specific 类的 interest rates for car loans 和 download grand theft auto.查询 presidents 的推荐结果中给出了 American presidents,university presidents 等结果,这些结果可被认为属于不同类别,但是,也可以认为各种 presidents 都是表示领导这一类语义概念.查询 interest rates for car loans 和 download grand theft auto 的推荐结果又被细分为多个类别,例如,interest rates for car loans 的结果中包含 car loan,mortgage rates,credit,refinance car 等类别.尽管 interest rates for car loans 和 download grand theft auto 这两个查询本身含义明确,但是搜索者可能还关心其他相关话题,Jigsoo 的分类结果也是有一定意义的.

Table 1 Classification results between vague and clear queries

表 1 TECH 用于辨别模糊查询和明确查询的分类结果

	Precision (%)	Recall (%)	*F*-Measure (%)
Vague query	85.71	92.31	88.89
Clear query	80.00	66.67	72.73

4.2 TECH能否识别模糊查询不同语义

前面分析了 TECH 对明确查询和模糊查询的区分能力,下面测试其对模糊查询不同语义概念的识别效果,这一节对比等式(2)和等式(4)对语义概念的识别能力,下一节对 Jigsoo 与 3 个对比系统进行比较.

首先介绍这两部分所用的测试查询.由于第 4.3 节中两个商业对比系统均为英文搜索引擎,所以我们只选用英文测试查询.实验中的标注者和被试由中国人组成,为了使大家的判断准确而不因对英语不熟而受影响,我们需要尽量选取为大家所熟知的英文歧义查询.首先,我们从英文维基百科上随机抽取 100 个歧义术语,这些术语的标题会被“(disambiguation)”标识.然后,我们从 Google Trends 和 Google zeitgeist 中抽取 50 个热门查询.这 150 个查询构成一个候选查询集,3 名同学分别浏览所有这些查询并标出他/她认识并确信其有歧义的查询,最终有 18 个查询获得了大家的一致认可.

实验基于智能查询推荐原型系统 Jigsoo,通过修改 Jigsoo 系统得到两个对比系统:(1) Baseline,将查询关系网络看作无权无向图,使用原始的模块度算法(公式(2))探测社团结构;(2) BL+TECH,采用修改后的算法(公式(4))探测社团结构.为了滤除其他因素的影响,构建词关系网络时没有利用 WordNet,这里只判断语义社团划分效果.3 名学生参与了此次评测,包括 2 名男生和 1 名女生,他们均为计算机研究生,且均熟知搜索引擎和信息检索,但是对本研究及该次实验的细节均不了解.每位评测者需要对每个实验系统的每个测试查询的返回结果进行如下 6 个方面的估计:

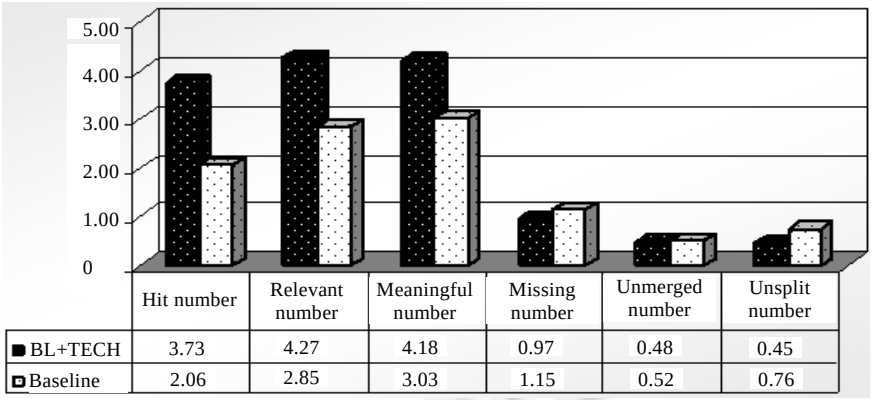
- Hit Number:聚类和标签都正确的类别个数;
- Relevant Number:聚类正确的类别个数;
- Meaningful Number:标签正确的类别个数;
- Missing Number:遗漏的与查询相关的其他含义个数;
- Unmerged Number:本应合并却没有合并的类别个数;
- Unsplit Number:本应分裂却没有分裂的类别个数.

各个评测者所面对的待评测数据相同,对每个查询在每个实验系统上的结果,上述各个数值被标注给出后,最终以 3 个标注的平均值作为最终结果.

对原始基于模块度的算法(公式(2))进行修改后的算法(公式(4))更适合查询相关词关系网络,对语义概念识

© 中国科学院软件研究所 http://www.jos.org.cn

别的准确率有很大的提高.图 4 给出了两种算法在识别类别数量上的对比结果.成对检验的单边 T 检验表明:TECH 在 Hit Number(p=2.02×10^{-8}),Relevant Number(p=7.8×10^{-5}),Meaningful Number(p=8.5×10^{-3})这 3 项指标上有显著提高;同时,在 Missing Number,Unmerged Number,Unsplit Number 这 3 项指标上的错误数均有所下降.



	Hit number	Relevant number	Meaningful number	Missing number	Unmerged number	Unsplit number
BL+TECH	3.73	4.27	4.18	0.97	0.48	0.45
Baseline	2.06	2.85	3.03	1.15	0.52	0.76

Fig.4 Comparison between TECH and original modularity algorithm

图 4 TECH 与原始模块度算法的比较

4.3 Jigsoo与其他系统的比较

为了评价 TECH 算法生成的智能查询推荐结果的有效性,我们设计了一个用户交互实验来比较 Jigsoo 和其他 3 个商业和学术系统产生的层次化目录.3 个对比系统为:两个商业搜索引擎 Cuil(http://www.cuil.com)和 Vivisimo(http://clusty.com)、一个基于文献[11]中图划分算法实现的智能查询推荐系统.

我们先简要说明一下这 3 个对比系统.Vivisimo 是一个知名查询结果聚类元搜索引擎,并多次被研究人员用作对比系统.尽管智能查询推荐不同于查询结果聚类,但是由于 Jigsoo 也是一种元搜索引擎,而且其生成的推荐目录也主要基于搜索引擎返回结果,这些都与 Vivisimo 技术手段类似.因此,这里选用 Vivisimo 作为对比系统.Cuil 由 Googlebase 前首席技术官等一批资深搜索引擎专家创立.它推出的查询推荐中包含一种层次化的查询推荐目录 Explore by Category.此外,我们还实现了一个基于图划分算法的智能查询推荐系统 StarClus,其采用的图划分算法为 star clustering[27].该算法曾被 Wang 等人[11]用于查询结果聚类并生成良好的类别目录.

为了尽量减少其他因素对实验评估结果的影响,我们对每个测试查询在每个系统上的返回结果做了统一化处理.由于 Jigsoo 基于 Google 的返回结果,而 Google 的返回结果随时都在变化,因此我们事先一次性地把每个测试查询对应的搜索返回结果下载到本地.这样,Jigsoo 和 StarClus 的结果都基于同一数据.用户在使用 Vivisimo 和 Cuil 搜索引擎网站时,有时候 Vivisimo 和 Cuil 初始化显示的目录结果只是部分结果,这可能是由于受浏览页面大小等原因造成的.为了使各个系统间公平地比较,我们在显示 4 个不同系统结果时,每个系统的所有结果都会直接向用户显示.另外,4 个不同系统的输出结果最后将在统一的界面上展示,并分别被标以 system1,system2,system3,system4,用户不知道每次每个系统对应的具体标号.经过这些操作,我们尽量减少了其他因素对用户评估结果的影响.

我们召集了 55 个被试参加本次交互实验,其中 39 人为男性,16 人为女性,且多数为学生,他们都熟知搜索引擎.在实验开始前,我们向他们讲述了本次实验的目的和内容,然后每个被试会被要求阅读一页具体的实验说明文件.说明文件包含查询推荐定义和智能查询推荐的界面和功能描述,但不包含任何主观倾向性内容.另外,说明文件还包含了对用户问题和实验要求的一些说明.

9 个查询用于用户交互实验:“bank rescue”,“Nobel”,“mouse”,“player”,“Beijing Olympic”,“China”,“apple”,“paper”,“ph”.如果不考虑被试之间的差异,实验中有两个独立变量:测试查询和对比系统.我们设计了 6 个问题来评价各个系统对各个测试查询生成的目录结果.测试查询有 9 个等级,对比系统有 4 个等级,这样可产生 36 种

© 中国科学院软件研究所 http://www.jos.org.cn

组合方式.为了获得更多的评价结果,我们采用了“within-groups”的实验设计方法[28].每个被试对随机分配 3 个查询的生成结果进行评价.对每个分配查询,4 个系统的结果向被试展示,各个系统的结果按随机顺序排列以排除学习迁移(transfer of learning effects)和位置因素对用户评价的影响.

对每个系统的生成目录,被试将从 6 个不同方面对其进行评价并回答 6 道选择题.其中 4 个问题关于是否应该添加、删除、合并、分裂目录中的类别,另外两个问题关于目录的全面性和条理性.分别以“添加”和“全面性”这两个问题为例:

(1) 对于该系统推荐出来的各类相关搜索,你是否会增加类别?

A. 不增加 B. 增加一到两个 C. 增加几个 D. 需要增加很多

(2) 你是否觉得该系统的推荐结果涵盖了该查询的几种主要含义?

A. 主要含义都涵盖了 B. 丢失了少部分主要含义 C. 丢失了很多主要含义 D. 结果完全无关

选项 A,B,C,D 分别会被量化为分值 1,2,3,4.这样,系统结果越好,分值就越低.其他问题也都以类似的方式组织,共形成 6 个评价指标:需添加的类别数(adding)、需删除的类别数(deleting)、需合并的类别数(merging)、需分裂的类别数(splitting)、丢失语义程度(coverage)、组织杂乱程度(tidiness).表 2 给出了各个系统在各个指标上得分平均值的比较,其中,“总平均值”是指 6 个指标平均值的总平均值.评估完 3 个查询后,每个被试需要完成一套关于使用各系统感受的问卷.

Table 2 Average responses about term hierarchy quality of different systems (lower=better)

表 2 用户对不同系统语义概念识别结果的平均评分(分值越低越好)

	Adding	Removing	Merging	Splitting	Coverage	Tidiness	All
StarClus	1.56	1.76	1.85	1.4	1.51	2.11	1.7
Jigsoo	1.59	**1.45**	**1.61**	1.47	1.33	**1.72**	**1.53**
Vivisimo	**1.08**	3	2.92	**1.21**	**1.28**	2.66	2.03
Cuil	1.8	1.7	1.84	1.4	1.7	2.02	1.74

4.3.1 什么才是好的查询推荐方式?

在用户实验中,每个被试都回答了这样一个问题:

如果现在由你来设计一个查询推荐系统,你希望以什么方式来展示相关搜索?

A. 像百度和 Google 那样将所有相关搜索放到一起

B. 根据相关搜索涉及的含义或领域将它们组织分类展示

C. 对含义明确查询采用选项 A 中方式展示,对歧义或概念广泛查询采用选项 B 中方式展示

D. 其他方式

我们分别以 List,Hierarchy,ITS(intelligent term suggestion),other 表示选项 A,B,C,D.图 5 给出了被试回答的统计结果.从图中我们可以看出,多数用户希望看到智能查询推荐方式,而且选择智能查询推荐的被试比例远大于选择其他方式的.

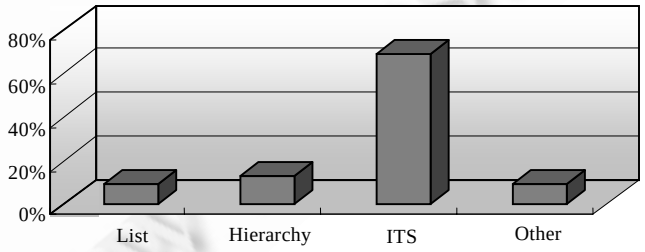


Fig.5 User preferences on the manner of term suggestion

图 5 喜好不同查询推荐方式的用户比例

© 中国科学院软件研究所 http://www.jos.org.cn

4.3.2 Jigsoo 与 Vivisimo,Cuil 的比较

从表 2 中我们可以看出,总体上 Jigsoo 优于 Vivisimo 和 Cuil.方差分析显示,在显著性水平 0.01 下,Jigsoo 在“需删除的类别数”、“需合并的类别数”、“组织杂乱程度”这些指标上优于其他系统,Jigsoo 在“丢失语义程度”上与 Vivisimo 表现差不多且好于 Cuil(Tukey's post-hoc tests: all p<0.05).Jigsoo 的总平均值最低,说明用户更接受 Jigsoo 的结果.尽管 Vivisimo 在“需添加的类别数”、“需分裂的类别数”上统计显著优于其他两个系统,但是其在“需删除的类别数”、“需合并的类别数”、“组织杂乱程度”上表现很糟,这导致其总平均值最差.

TECH 算法与图划分算法的比较.TECH 算法在模糊查询语义概念识别上优于图划分算法 StarClus.使用与构建 Jigsoo 同样的方法,我们搭建了基于图划分算法“star clustering”[27]的智能查询推荐系统 StarClus.从表 2 中可以看出,Jigsoo 在多数指标上优于 StarClus 或与其不相上下.T 检验表明,Jigsoo 和 StarClus 在“需添加的类别数”和“需分裂的类别数”上没有显著差异,而在其他指标上,Jigsoo 明显优于 StarClus(p<0.01).

4.3.3 被试对各个系统的综合评价

在最后的调查问卷中,每个被试对各个对比系统的可用性作做了一个综合评估.首先是系统可用性测试,对每个系统 *X*,被试要回答下列问题:

在实际搜索中,你是否会采用系统 *X* 的推荐查询?

A. 肯定会使用 B. 会使用 C. 不会使用 D. 绝不会使用

这些选项也会分别被量化为 1,2,3,4 分,然后进行统计分析.最后,每个被试要回答以下问题:

上述 4 个系统中,你最喜欢哪个系统?

A. 系统 1 B. 系统 2 C. 系统 3 D. 系统 4

表 3 和图 6 分别给出了这两个问题的统计结果,Jigsoo 在可用性和用户喜好度上均优于其他系统.在可用性测试中,Jigsoo 是唯一均值小于 2 的系统,T 检验显示,Jigsoo 显著优于其他系统(p<0.05).

Table 3 Responses to the question: Would you use this system for term suggestion in practice?

表 3 对问题“在实际搜索中,你是否会采用系统 X 的推荐查询?”的回答情况

	StarClus	Jigsoo	Vivisimo	Cuil
Definitely Yes	4	**9**	2	8
Yes	34	**38**	20	26
No	16	**8**	25	19
Definitely No	1	**0**	8	2

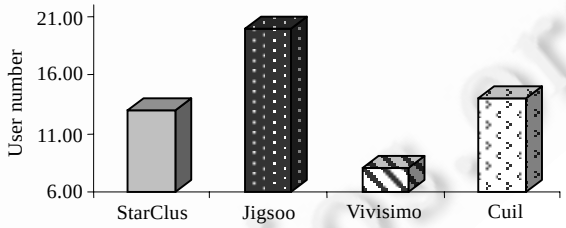


Fig.6 User preferences on each system

图 6 喜欢各个系统的被试数

5 总结和未来工作

信息检索的目标在于满足用户的信息需求,现代信息检索技术要求用户先将其信息需求转化成一个简短的查询,然后信息检索系统针对这个查询给予匹配信息.查询是用户及其所需信息之间的桥梁.如何帮助用户构造一个能正确表达其信息需求的合适查询是信息检索中的关键问题之一.针对这一问题,本文根据自然语言的小世界性质提出一种智能查询推荐技术.智能查询推荐可以自动辨识一个查询是否含义明确.如果查询含义模糊,则智能查询推荐将根据该查询的各种可能意图构建一个由相关查询组成的语义概念目录.为了实现智能查

© 中国科学院软件研究所 http://www.jos.org.cn

询推荐,我们综合利用了物理学复杂网络中的社区发现算法和信息检索技术,提出一种查询词语义识别算法TECH.基于TECH,我们搭建了一个简单的智能查询推荐原型系统Jigsoo.为了正确评价智能查询推荐的可用性和TECH算法的效果,我们设计了评价方法和准则,并利用Jigsoo做了一系列实验.实验结果表明:相对于传统的查询推荐方式,用户更喜欢智能查询推荐;基于TECH搭建的Jigsoo能够有效辨别模糊查询;Jigsoo构建的语义概念分类目录优于其他3个商业和学术对比系统.

在初步实验中,Jigsoo已经表现出了令人振奋的结果.我们相信,在智能查询推荐模式和TECH算法框架下,通过引入更多的资源和更有效的信息检索算法,用户与搜索引擎的交互将会变得更加容易.在未来的工作中,我们还计划开展以下研究:

- 搜索引擎查询日志是当前查询推荐中用到的一种主要资源,结合基于查询日志的查询推荐方法,利用TECH对查询日志进行处理,可能会得到更好的智能推荐结果.
- Jigsoo中相关查询词的选取、词关系网络的构造都采用了一些很简单的方法,以后需要考虑融合进关键词抽取、数据挖掘中一些更有效的算法来提高智能查询推荐的效果.
- 由于缺乏合适的查询评价集和对比系统,本文没有对中文智能查询推荐的效果进行分析评价.尽管中文也符合自然语言的小世界性质,但也可能有一些特殊性质,我们计划下一步开展中文方面的智能查询推荐研究.
- 本文的工作在很多地方得益于物理学家对自然语言和真实复杂网络的研究,物理学相关领域的最新进展也将会对智能查询推荐技术的完善起到有益的帮助.

References:

[1] Song R, Luo Z, Wen JR, Yu Y, Hon HW. Identifying ambiguous queries in Web search. In: Proc. of the 16th Int'l World Wide Web Conf. (WWW 2007). New York: ACM, 2007. 1169−1170. [doi: 10.1145/1242572.1242749]

[2] Huang CK, Chien LF, Oyang YJ, Relevant term suggestion in interactive Web search based on contextual information in query session logs. Journal of the American Society for Information Science and Technology, 2003,54(7):638−649. [doi: 10.1002/asi.10256]

[3] Wang JM, Peng B. User behavior analysis for a large-scale search engine. Journal of the China Society for Scientific and Technical Information, 2006,25(2):154−162 (in Chinese with English abstract).

[4] Cui H, Wen JR, Li MQ. A statistical query expansion model based on query Logs. Journal of Software, 2003,14(9):1593−1599 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/14/1593.htm

[5] Xu J, Croft WB. Query expansion using local and global document analysis. In: Proc. of the ACM-SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM, 1996. 4−11. [doi: 10.1145/243199.243202]

[6] Jones R, Rey B, Madani O, Greiner W. Generating query substitutions. In: Proc. of the 15th Int'l Conf. on World Wide Web (WWW 2006). New York: ACM, 2006. 387−396. [doi: 10.1145/1135777.1135835]

[7] Belkin NJ. Helping people find what they don't know. Communication of ACM (CACM), 2000,43(8):58−61. [doi: 10.1145/345124.345143]

[8] Hearst MA, Pedersen JO. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In: Proc. of the 19th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM, 1996. 76−84. [doi: 10.1145/243199.243216]

[9] Joho H, Sanderson M, Beaulieu M. Hierarchical approach to term suggestion device. In: Proc. of the 25th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2002). New York: ACM, 2002. 454−461. [doi: 10.1145/564376.564495]

[10] Ferragina P, Gulli A. A personalized search engine based on Web-snippet hierarchical clustering. In: Special Interest Tracks and Posters of the 14th Int'l Conf. on World Wide Web. New York: ACM, 2005. 801−810. [doi: 10.1145/1062745. 1062760]

[11] Wang XH, Zhai CX. Learn from Web search logs to organize search results. In: Proc. of the 30th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM, 2007. 87−94. [doi: 10.1145/1277741.1277759]

[12] Stoica E, Hearst M, Richardson M. Automating creation of hierarchical faceted metadata structures. In: Proc. of the Human Language Technologies: The Annual Conf. of the North American Chapter of the Association for Computation Linguistics (NAACL-HLT 2007). Rochester, 2007. 244−251.

[13] Cancho RFI, Solé RV. The small world of human language. Proc. of the Royal Society B: Biological Sciences, 2001,268:

© 中国科学院软件研究所 http://www.jos.org.cn

2261−2265. [doi: 10.1098/rspb.2001.1800]
[14] Liu ZY, Sun MS. Chinese word co-occurrence network: Its small world effect and scale-free property. Journal of Chinese Information Processing, 2007,21(6):52−58 (in Chinese with English abstract).
[15] Li YN, Zhang S, Wang B, Li JT. Characteristics of Chinese Web searching: A large-scale analysis of Chinese query logs. Journal of Computational Information Systems, 2008,4(3):1127−1136.
[16] Yu HJ, Liu YQ, Zhang M, Ru LY, Ma SP. Research in search engine user behavior based on log analysis. In: Proc. of the SWCL 2006. 2006. 76−80 (in Chinese with English abstract).
[17] Watts DJ, Stogatz SH. Collective dynamics of 'small-world' networks. Nature, 1998,393(6684):440−442. [doi: 10.1038/30918]
[18] Cheng XQ. Analysis the topological structure and the content relevance of the information networks [Ph.D. Thesis]. Beijing: Institute of Computing Technology, the Chinese Academy of Sciences, 2005 (in Chinese with English abstract).
[19] Newman MEJ. Analysis of weighted networks. Physical Review, E, 2004,70:056131. [doi: 10.1103/PhysRevE.70.056131]
[20] Shi JB, Malik J. Normalized cuts and image segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), 2000,22(8):888−905. [doi: 10.1109/34.868688]
[21] Zhang D, Mao R. Classifying networked entities with modularity kernels. In: Proc. of the 17th ACM Conf. on Information and Knowledge Management (CIKM 2008). New York: ACM, 2008. 113−121. [doi: 10.1145/1458082.1458100]
[22] Wang XF, Li X, Chen GR. Complicated Network Theory and Application. Beijing: Tsinghua University Press, 2006.
[23] Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review, E, 2004,69:026113. [doi: 10.1103/PhysRevE.69.026113]
[24] Newman MEJ. Finding community structure in networks using the eigenvectors of matrices. Physical Review, E, 2006,74:036104. [doi: 10.1103/PhysRevE.74.036104]
[25] Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Physics Review, E, 2004,70:066111. [doi: 10.1103/PhysRevE.70.066111]
[26] Nguyen VB, Kan MY. Functional faceted Web query analysis. In: Proc. of the 16th Int'l World Wide Web Conf. (WWW2007) Workshop on Query Log Analysis: Social and Technological Challenges. New York: ACM, 2007. http://www2007.org/workshop-W6.php
[27] Aslam JA, Pelekhov E, Rus D. The star clustering algorithm for static and dynamic information organization. Journal of Graph Algorithms and Applications, 2004,8(1):95−129.
[28] Dix AJ, Finlay JE, Abowd GD, Beale RB, Finley JE. Human-Computer Interaction. 2nd ed., Prentice Hall, Inc., 1998.

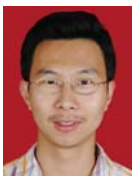
附中文参考文献:

[3] 王继民,彭波.搜索引擎用户点击行为分析.情报学报,2006,25(2):154−162.
[4] 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. http://www.jos.org.cn/1000-9825/14/1593.htm
[14] 刘知远,孙茂松.汉语词同现网络的小世界效应和无标度特性.中文信息学报,2007,21(6):52−58.
[16] 余慧佳,刘奕群,张敏,茹立云,马少平.基于大规模日志分析的网络搜索引擎用户行为研究.见:第 3 届学生计算语言学研讨(SWCL 2006).2006. 76−80.
[18] 程学旗.信息网络拓扑结构与内容相关性研究[博士学位论文].北京:中国科学院计算技术研究所,2005.
[22] 汪小帆,李翔,陈关荣.复杂网络理论及其应用.北京:清华大学出版社,2006.

李亚楠(1984－),男,山东东营人,博士,主要研究领域为信息检索.

李锦涛(1962－),男,博士,研究员,博士生导师,主要研究领域为数字媒体处理技术,虚拟现实技术,普适计算技术.

王斌(1972－),男,博士,副研究员,博士生导师,主要研究领域为信息检索.



李鹏(1985－),男,博士生,主要研究领域为信息检索.

© 中国科学院软件研究所 http://www.jos.org.cn