基于规则和类型还原的用户查询意图识别*

王俞霖[1,2] 孙乐[1] 黄云平[1,2] 李文波[1]
1、中国科学院软件研究所 北京 100190
2、中国科学院研究生院 北京 100049

摘 要：识别网络查询隐含的用户意图是一项具有重要意义和挑战性的工作。本文通过对真实用户查询日志的标注和分析，发现基于规则的方法可以对用户意图进行有效的识别。针对信息类、导航类和事务类三种用户意图，我们总结出若干规则对其进行自动识别。之后我们提出一种类型还原方法可以进一步提高查询意图识别的召回率。实验的结果表明，基于规则和类型还原相结合的方法能够有效的对用户意图进行识别。

关键词：规则方法；类型还原；查询意图识别

User Query Intent Identification Based on Rules and Type Restoration

WANG Yu-lin[1,2], SUN Le[1], HUANG Yun-ping[1,2], LI Wen-bo[1]
1 Institute of Software, Chinese Academy of Science, Beijing 10080
2 Graduate University of Chinese Academy of Science, Beijing 10049

Abstract: Identifying the intent hidden in user queries is a task of great importance and challenge. By analyzing and manually labeling a large number of real user query log, we found rules were very useful in identifying user intents. A batch of rules was constructed to identify user intent automatically. Then we proposed a method called type restoration to improve the recall rate of identification. The experimental results show that the method of rules and type restoration is effective and efficient in user intent identification.

Key words: rules; type restoration; query intent identification

1 引言

网络查询中隐式地包含用户意图。所谓用户意图，指用户向搜索引擎提交查询的动机。研究表明，用户意图大体可以分为导航类、信息类和资源类三种。根据用户意图不同采用最合适的排序算法可以获得更好的搜索结果，所以识别用户意图具有重要的意义。

通过对用户查询意图的识别来提高搜索的质量，需要经过三个基本的步骤：首先需要确定用户意图分类的框架；第二需要对查询意图进行自动识别；最后需要根据不同的用户意图给出更优的结果。本文在前人工作的基础上，提出一套适应于中文查询的用户意图的二层分类框架，通过对 12,500 条真实查询的手工标注，总结出若干条实用规则用于对查询意图进行自动识别。使规则方法的缺点是召回率偏低。为解决该问题我们提出类型还原的方法用于提高用户查询意图识别的召回率。

2 相关工作

* 本文相关研究得到国家自然科学基金资助重点项目(60736044),国家 863 计划重点项目(2006AA010108-5),国家 863 计划资助项目(2008AA01Z145),国家自然科学基金资助项目(60773027)资助。

传统的信息检索理论认为用户查询意图都是为了获得某种信息。Broder[1]最早将查询意图扩展为三大类：导航类，信息类和事务类。Rose[2]等对 Broder 分类体系进行了修改和扩充，用资源类代替事务类意图，构造了一个层次式的用户意图分类框架，并采用人工的方法对网络查询日志进行分类。

Kang[3]等提出使用查询词分布，互信息，链接文本使用率和词性信息的方法对查询意图进行自动识别。Baeza-Yates[4]按照信息类、非信息类的分类框架对用户查询意图进行自动的分类，采用监督和非监督方法相结合的办法得到较好的结果。Lee[5]的研究表明，用户点击信息和查询词在链接锚文本中的分布是对查询意图进行识别的有效特征。Fujii[6]使用改进了的查询词在锚文本中的分布的方法对用户意图进行识别。以上的工作都回避了对事务类查询进行识别，采用的数据多数不是真实的搜索引擎查询。

Jansen[7]等总结出 20 条规则用于对用户意图进行识别，实验结果正确率达到 74%。张森[8]采用点击偏好的方法区分导航类和非导航类，然后利用动词信息对资源类和信息类的查询进行区分，在真实查询数据集上取得了良好的结果。

上述可知，前人的工作采用的用户意图分类框架和查询数据并不统一，而且采用的语料也不相同。本文中，我们采用的分类框架遵循主流的 Broder 框架，并针对中文查询进行稍微的改动，实验的数据来自百度搜索引擎公布的用户查询日志。

3 用户意图分类框架的确定和手工标注

受 Rose 的层次性用户意图分类的启发，我们根据汉语查询的特点为查询中隐含的用户意图构建一个二层的分类框架。

实验中使用的查询来自百度瞬时风向标。我们选取共计 12,500 条查询进行处理和分析。首先我们使用导航类，信息类和资源类作为用户意图分类的第一层框架。随后针对收集来的用户查询确定第二层的分类框架。进行二级分类框架有以下几个原则：类别应该有代表性，即查询集中应该有存在大量的查询属于该类别；类别本身具有重要的意义；类别应该不和其它已存在的类别重叠；所有的类别应该能覆盖绝大多数查询。最终确定的用户意图框架如表 1 所示：

表 1 用户意图分类框架

用户意图		描述	例子
Navigational	N	网站	新浪 搜狐 土豆 豆瓣
Informational	IB(Buy)	欲购买的商品信息	手机 汽车 笔记本 数码相机
	IE(Event)	近期发生的重大事件信息	北京奥运 微软黑屏 汶川地震
	IA(Advice)	征求做某事的建议	如何戒烟
	II(Inquiry)	查询结果	基金净值 手机归属地 天气预报
Resource	RE(Experience)	体验类资源	电影 电视剧 音乐 小说 漫画
	RA(Assistant)	辅助类资源	试题 菜谱 范文 地图 歌词
	RI(Install)	安装类资源	软件 游戏 驱动程序
UN		无法预测意图的查询	3.1 2k9 长江

我们对收集来的 12,500 条查询，按照上面的分类框架进行手工标注。文献[2]中使用四种信息来源对用户的意图进行识别，分别是查询本身，搜索引擎返回的结果，用户点击信息和用户的后继查询。文献[2]指出，只用查询本身所得到的标注结果和使用四种信息所得到的标注结果的效果几乎是一样的，所以我们在标注查询意图的过程中也只是使用查询本身作为唯一的评判标准。最终得到的结果如图 1 所示：

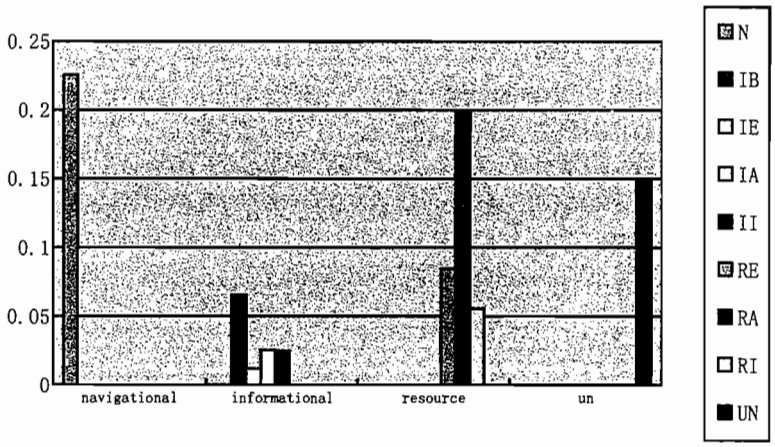

图 1 各种意图查询分布图

如图 1 所示，不能确定用户意图的查询的比例为 14.83%，考虑到用户输入中含有噪声，该框架的覆盖率的效果比较理想。另外值得注意的是，关于导航类，信息类和资源类三者各占的比例中，图 1 中资源类超过了信息类，这与前人的结论有所不同。

4 自动分类方法

本文使用规则的方法对用户意图进行分类。规则的方法实现简单，分类的效果很好。其次规则的方法具有通用性。用户输入的查询的具体内容总是在不断变化的，但是以规则表现出的内部结构不会有太多的变化。而且文献[7]使用规则的方法适用于大规模的数据量，这是其他方法所不具备的。

4.1 规则的构造

通过对查询日志的分析，得到用于识别用户意图的规则如下：

Navigational：含有 url 的组成部分；含有表示其为网站的词；含网站名常用词；含有地名和组织机构名等。Informational：含有表示疑问的词；含有口语词；含有表示信息的词；含有商品名，厂商名和型号；含有表示事件的词；含有服务名；查询由两个短语组成且之间可以插入“的”；含有表示指导的动词，名词和助词；含有常见的表示查找结果的词；含有人名或团队名等。Resource：查询中含有“下载”；查询中含有表示辅助类资源的词语；查询中含有表示体验类资源的词；查询中含有与软件、游戏和驱动等表示安装类资源相关的词语；查询中含有表示交互的词；查询中含有表示多媒体文件格式的后缀等。

使用基于规则的方法对用户的意图进行识别，存在两个问题。第一，对于查询所有的规则都不能和它匹配；第二，有超过一条的规则和查询匹配。对于第二种情况，经过分析发现，规则匹配的位置越靠后，它的重要性往往越大。所以，当查询与多条规则同时匹配的时候，我们保留匹配位置最靠后的一条。

4.2 类型还原的定义及使用

经常出现查询与所有的规则都不匹配的情况，造成对用户意图识别的召回率偏低。互联网

更新很快，我们把新出现的代表某事物的词语称为实体词。如果能把实体词还原为规则中已有的类型，那么必将大大提高用户意图的识别率。

我们将实体定义为一个由五类信息组成的五元组，分别是实体类型、实体动作、实体名、属性和实体独一无二性词语，表示为 M:{T, E, A, P, U}。其中，实体类型信息对识别用户意图最有用，但它并不总出现在查询中。如果能够通过其它信息推测出当前的实体的类型，那么对识别该查询的意图提供很大帮助。这种方法称之为类型还原。

为了能够推测出不完整查询的类型，需要获得实体模型中的其他信息。根据手工标注和简单规则识别的结果，本文选取了 12 种实体类型，每种实体类型对应一种用户意图。百度提供的 Query Suggestion 功能是获得一个查询实体模型中缺失部分的重要来源。为了建立模型中缺失的 A、P 和 U 部分，本文从 12 种实体类型中分别选取 50 个具有代表性的实体词，将它们提交到百度的 Query Suggestion，对返回的短文档中进行手工标注，将词语按照 A、P 和 U 的类别进行归类整理。其中 U 的构造过程比较特殊，因为它没有固定的格式。本文选用了 6 种最直观的独一无二性词语，采用正则表达式的方法对独一无二性词语进行的提取。

表 2 独一性词语表

代号	描述	示例
Episode	电视剧的剧集	妻子的诱惑第 34 集　越狱第四季　珠光宝气 46 集
Stock	股票代号	600029　600628
Year	年份	qq2008　word2003　cad2004 飞信 2008
Single	单数字版本号	暗黑 3　魔力宝贝 2　帝国时代 3
Fraction	小数版本号	卡巴斯基 8.0 dota6.59
Model	商品型号	奥迪 r8　5610　三星 i908　x61

为了判断实体词 e 所属的类型，我们采用公式(1)进行计算。

$$T(e) = \arg\max_t (Score_M(t, A', P', U')) \tag{1}$$

T(e)表示实体词 e 所属的类型，A'，P'和 U'是把 e 提交到百度 Suggestion 后所得到的对应于动作、属性和唯一性的词语的集合。M 代表已经建立好的实体词的模型。我们把实体词 e 划分到使得分函数最大的类型中去。得分公式见公式(2)。

$$Score_M(t, A', P', U') = \alpha \cdot Sub_M(t, A') + \beta \cdot Sub_M(t, P') + \gamma \cdot Sub_M(t, U') \tag{2}$$

得分公式由三个子得分公式组成，代表由动作词，属性词和唯一性词分别得到的分数。公式中的 α，β 和 γ 表示三个子公式的权重。子得分公式的定义由公式(3)给出。

$$Sub_M(t, A') = \sum_{a \in A'} tf_{t,M}(a) \cdot idf_M(a) \tag{3}$$

$tf_{t,M}(a)$表示词 a 在模型 M 中类型 t 的短文档（百度 Suggestion 返回的结果）中的词频，$idf_M(a)$表示词 a 在模型 M 中所有类型的短文档的逆文档频率。另外两个子得分公式的定义与之类似。得分公式中权重的定义见公式(4)。

$$\alpha = H(T) - H(T \mid A), \quad \beta = H(T) - H(T \mid P), \quad \gamma = H(T) - H(T \mid U) \tag{4}$$

H(T|A)表示在给定动作词时，实体类型的条件熵。直观上讲，信息增益越大，该种类的词语对实体类型区分的帮助越大。

5 实验

我们使用2008年12月的百度风向标查询作为测试数据，将3,000条查询平均分为3个测试集，每个测试集含有1,000条查询。实验结果评价的标准是正确率、召回率和F1值。

5.1 只使用规则的方法的实验结果

首先只使用简单的规则的方法进行测试，作为实验的baseline。实验结果如表3所示。

表3 只用规则方法实验结果

	Test Data Set 1			Test Data Set 2			Test Data Set 3		
	P	R	F1	P	R	F1	P	R	F1
N	0.913	0.660	0.766	0.867	0.650	0.743	0.927	0.651	0.765
I	0.796	0.195	0.314	0.797	0.213	0.336	0.783	0.221	0.345
R	0.906	0.413	0.568	0.883	0.434	0.582	0.916	0.427	0.583

实验的结果表明，使用规则的方法能够获得很高的准确率，但是召回率偏低。尤其是对信息类意图，召回率只有20%左右，这是由信息类的查询形式多变，难以构造有效的规则所导致的。而对于导航类意图，规则方法能够有效地识别，原因是导航类的查询格式相对固定。资源类的结果处于导航类和信息类之间。实验结果表明信息类查询意图最难识别。

5.2 使用规则和类型还原方法的实验结果

在此基础上我们使用4.2节中提出的类型还原的方法再次进行实验，结果如表4所示。

表4 使用类型还原方法的实验结果

	Test Data Set 1			Test Data Set 2			Test Data Set 3		
	P	R	F1	P	R	F1	P	R	F1
N	0.795	0.771	0.783	0.820	0.742	0.779	0.873	0.748	0.805
I	0.759	0.636	0.692	0.761	0.593	0.667	0.668	0.599	0.632
R	0.851	0.691	0.763	0.867	0.718	0.786	0.894	0.729	0.803

使用类型还原后，结果中仍表现出准确率高于召回率的现象，但是各项召回率均有不同程度的提高。对于最难的信息类，F1值接近70%，提高了将近一倍。实验结果表明，类型还原的方法能够有效地对用户查询的意图进行识别。

5.3 失败样例的分析

我们对意图识别失败的查询进行分析和归类。意图识别失败的原因大体可以分为4类：

- 规则方法的误判，如“坏蛋是怎样炼成的”中，含有“怎样”被规则方法判定为IA(advice)类，实际上它是一部小说的名字。这类错误约占错误总量的6.8%。
- 规则库建立的不完善，如“小路工作室”，“浩方优化版”等。如果规则库建立的完备，应该能够正确识别。这类错误约占错误总量的14.9%。
- 类型还原出错，如“李维斯”被还原成网站，其实是一个品牌的名字。这类错误约占错误总量的7.2%。
- 规则和类型还原都没有能够识别出用户的意图，这类错误约占错误总量的66.1%。

5.4 实验结果与前人的比较

为了对本文基于规则和类型还原方法的性能进行评价，我们将得到实验的结果与文献[7]以及文献[8]的实验结果进行对比。文献[7]、文献[8]和本文都使用 Broder 的分类框架对用户意图进行识别，而且分类的对象均为来自搜索引擎查询日志，具有很好的可比性。

文献[7]得到的分类正确率为 74%，而本文的方法对导航类，信息类和资源类查询得到的正确率的中位值分别为 82.0%，75.9%和 86.7%。考虑到信息型意图难以识别，本文方法的准确率超过了 Jansen。而召回率及 F1 值在 Jansen 的文章里没有提及。文献[8]对导航类的 F1 值为 88%，资源类(或事务类)的 F1 值为 85%，而本文的方法得到这两项中位值分别为 78.3%和 78.6%，低于文献[8]的实验结果。文献[8]所使用的测试集数据量较小，且不含噪声。本文使用的数据集并没有手工筛选，其中含有约 15%的噪声。实际查询中含有噪声是不可避免的。本文的方法在实际中可用性应该会好一些。

6 结论及展望

本文通过对 12,500 条真实搜索引擎的手工标注，构造了一套适用于中文查询意图的二级分类框架。在此分类框架基础上，我们采用规则的方法对用户的意图进行自动识别。为了解决规则方法召回率偏低的问题，我们提出类型还原的方法，提高召回率。通过对实验结果的分析发现，相对于导航类和资源类，信息类意图的识别是最困难的。随后我们对识别失败的样例进行分析，最后将本文的实验结果与前人的工作进行对比，表明本文基于规则和类型还原的方法能够有效地识别用户查询意图。

识别用户查询意图，可以针对不同的意图采用不同的排序算法，或者给出不同的结果布局，使用户获得更好的检索结果。这是我们下一步工作的内容。

参 考 文 献

[1] Andrei Broder. A taxonomy of web search. ACM SIGIRForum. 2002, 3-10.

[2] Daniel E. Rose, Danny Levinson. Understanding user goals in web search. Proceedings of the 13th international conference on World Wide Web. 2004, 13-19.

[3] In-Ho Kang, GilChang Kim. Query type classification for web document retrieval. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. 2003, 64-71.

[4] Ricardo Baeza-Yates, Liliana Calderón-Benavides, Cristina González-Caro. The Intention Behind Web Queries. Lecture Notes in Computer Science, 2006, Volume 4209/2006: 98-109.

[5] Uichin Lee, Zhenyu Liu, Junghoo Cho. Automatic identification of user goals in Web search. Proceedings of the 14th international conference on World Wide Web. 2005, 391-400

[6] Atsushi Fujii. Modeling Anchor Text and Classifying Queries to Enhance Web Document Retrieval [A]. Proceedings of the 17th international conference on World Wide Web. 2008, 337-346

[7] Bernard J. Jansen, Danielle L. Booth, Amanda Spink. Determining the user intent of web search engine queries. Proceedings of the 16th international conference on World Wide Web. 2007, 1149-1150.

[8] 张森. WEB 检索查询的意图分类研究. 中科院研究生院硕士学位论文. 2008.