КЛАССИФИКАЦИЯ ОСНОВНЫХ ТИПОВ ЗАПРОСОВ ПОЛЬЗОВАТЕЛЯ В СИСТЕМЕ ИНФОРМАЦИОННОЙ ПОДДЕРЖКИ ИННОВАЦИОННОЙ ДЕЯТЕЛЬНОСТИ

Д. Ю. Постаногов

Белорусский государственный университет
Минск, Беларусь
E-mail: denpost@gmail.com

В данной работе рассматривается проблема классификации запросов в информационно-поисковых и вопросно-ответных системах, а также недостатки существующих подходов к классификации. Предложены классы основных типов запросов пользователя по выраженной в них информационной потребности для их применения в системе информационной поддержки инновационной деятельности.

Ключевые слова: классификация запросов, вопросно-ответные системы, поисковые системы, инженерия знаний.

Эффективность инновационной деятельности как творческого процесса, направленного на преобразование знаний в новые продукты, услуги или технологии, в большой степени определяется уровнем развития ее информационной поддержки, к

основным задачам которой можно отнести обеспечение изобретателя необходимыми знаниями на различных этапах инновационного процесса, в том числе на этапе фундаментальных исследований, прикладных исследований и разработок, опытно-конструкторских работ, а также при организации промышленного внедрения изобретений. Класс приложений, реализующих вопросно-ответные системы, является одним из наиболее востребованных инструментариев доступа к атрибутивным знаниям в аспекте указанной проблемы. В свою очередь, задача классификации запросов пользователя является важнейшей составляющей в разработке таких систем [1], при этом, однако, ее решение может преследовать различные цели в зависимости от используемого подхода.

Так, в системах информационного поиска наибольшее распространение получила категоризация запроса с целью вычисления ее тематической категории из заранее определенного списка (например, «sports», «shopping», «places» и т. п.) [2], которая, будучи принятой во внимание в алгоритме поиска по ключевым словам, позволяет повысить эффективность поисковой системы в целом за счет выбора результатов в соответствии с определенной темой. Очевидно, что данный подход не опирается на распознавание типов атрибутивных знаний и, таким образом, не обеспечивает возможность эффективного доступа к ним в рамках рассматриваемой проблемы.

В вопросно-ответных системах задача классификации запросов имеет более узкую направленность и заключается в определении семантических типов ответов, запрашиваемых в вопросе пользователя, выраженном на естественном языке (ЕЯ), при этом список возможных классов, в свою очередь, может иметь различную степень детализации. Так, например, в [3] рассматривается 13 общих категорий (Causal antecedent, Goal orientation, Ennoblement, Causal consequent, Verification, Disjunctive, Instrumental/procedural, Concept completion, Exceptional, Judgmental, Quantification, Feature specification, Request), накладывающих ограничение на вид ответа в вопросно-ответной системе. В [4] описывается решение проблемы получения одного или нескольких так называемых «Qtargets», соответствующих вопросу пользователя, из списка 122 возможных типов ответов (например, «temporal-quantity» – «How long would it take to get to Mars?», «proper-organization» – «Who made the first airplane?»). Существенным ограничением в данном подходе является то, что он ориентирован, в первую очередь, на поддержку вопросов так называемого "фактоидного" типа [5, 6], т. е. запрашивающих однозначные ответы в форме простых именованных сущностей, таких как дата, расстояние, размер, имя, организация и т. п., наличие которых в ответах обеспечивается дополнительным этапом «просеивания» результатов традиционного алгоритма поиска по ключевым словам.

Очевидно, что создание вопросно-ответной системы, способной точно и исчерпывающе отвечать на любой запрос пользователя на основании автоматического анализа большого объема текстовых документов различных областей знания, потребует значительных усилий в силу свойств ЕЯ как объекта моделирования, препятствующих безошибочному автоматическому анализу произвольного текста. Кроме того, поддержка таких запросов, как, например, «what are known nicknames of Pamela Anderson?», «give me a list of top 5 most Oscar-awarded Hollywood actors?», может потребовать не только введения специальных типов атрибутивных знаний, но также отдельных механизмов поиска и формирования списка ответов на основе логического вывода. Однако, как показали проведенные исследования, для обеспечения достаточно эффективной информационной поддержки пользователя в инновационной деятельности, необходимо учитывать наиболее востребованные в данной предметной

области типы атрибутивных знаний, тем не менее не ограниченные лишь фактоидной формой, при этом поддержка доступа к знаниям остальных типов может быть достигнута путем использования общих средств такой системы.

В силу этого разработка перечня основных типов атрибутивных знаний, наиболее актуальных для инновационной деятельности, была произведена путем анализа выборки 60 000 уникальных англоязычных ЕЯ-запросов пользователей системы автоматизации инженерии знаний и решения инновационных задач Goldfire (http://goldfire.com). При этом в качестве критериев для отбора того или иного типа атрибутивных знаний учитывались следующие факторы: возможность сведения различных запросов в общий класс выраженной в них информационной потребности; частота встречаемости класса в выборке, отражающая его востребованность в рамках рассматриваемой задачи; возможность извлечения знаний с достаточной полнотой и точностью общими механизмами без введения специального класса.

Так, например, запросы «what is the average pressure in combustion chamber?» и «what is the acceptable vibration amplitude of drill?» можно рассматривать в качестве принадлежащих одному классу информационной потребности, получившему формальное обозначение «QT_ParameterValue» и соответствующему атрибутивным знаниям о значениях различных физических параметров. Однако для отдельных параметров, таких как «size», «length», «width», «thickness», «depth», «speed» и т. д. такое обобщение является недостаточным в силу того, что запросы, направленные на извлечение знаний о значениях данных параметров, могут не содержать в явном виде их имен. Вместо этого в запросах (равно как и в текстовых документах) могут быть использованы, например, имена прилагательные: «how big is…», «how long is…», «how wide is…» и т. д., при этом обязательным является упоминание самого объекта, «носителя» параметра. В силу этого были введены специальные типы атрибутивных знаний QT_Size, QT_Length, QT_Width и т. д., соответствующие атрибутивным знаниям о значениях конкретных параметров самих объектов.

В результате проведенного анализа были выделены следующие основные классы запросов по выраженной в них информационной потребности (таблица).

Классификация основных типов запросов пользователя

№ п/п	Класс информационной потребности (формальное имя класса)	Примеры запросов, сводящихся к указанному классу информационной потребности
1	Определение объекта (QT_Definition)	«what is a ring laser gyroscope?» «definition of a hydrocarbon»
2	Значение параметра объекта (QT_ParameterValue)	«what is the weight of oxygen» «red blood cells temperature» «what is lifetime of catalyst?» «pressure of the combustion chamber»
3	Список параметров объекта (QT_Parameter)	«parameters of laser radiation»
4	Количественные характеристики объекта или действия (QT_Quantity, QT_Amount)	«How many blood vessels are in human» «amount of inductive power in contactless mode»
5	Местоположение объекта, место действия процесса (QT_Location)	«reflector mounting location» «where is the camshaft located?» «Where is ISCC power plants built?»
6	Составные части объекта (QT_Part)	«what are the parts of gyroscope?» «what comprise the bearing

7	Составные части объекта (QT_PartOf)	«what contains fermentation broth» «what are the components of electric drives?»
8	Материалы (QT_MadeOf)	«what are biopolymers made of» «what materials are used in implants?»
9	Физические свойства материала (QT_MaterialProperty)	«what are the properties of resin?»
10	Материалы, удовлетворяющие заданному свойству (QT_PropertyMaterial)	«what materials are fire-resistant?»
11	Функции объекта (QT_Function)	«What does resveratrol do» «what can Platelet Rich Plasma do»
12	Конкретизация (уточнение видов) объектов (QT_Instantiation)	«what types of long linear bearings are on market today?» «what are kinds of sol gel?» « types of porous polymer material»
13	Применение объекта (QT_Application)	«application of image sensor» «what is the application of H-peroxide?» «where is chitosan used in medical devices?» «ion exchange resin use»
14	Недостатки (QT_Disadvantage)	«disadvantages of solar cell» «what are the disadvantages of the faucet adapter?»
15	Достоинства (QT_Advantage)	«advantage of stirling engine» «What is the advantage of perspiration»
16	Объекты (классы объектов), удовлетворяющие заданным условиям (QT_Object)	«what can be separated from hydrogen» «what can LCD combine with?» «Which materials are viscoelastic» «what animals may be used to test steroids?»
17	Инструменты достижения эффекта, осуществления процесса (QT_Subject)	«What can detect biological agents?» «what can affect lactame polymerization rate» «which types of monitors are less expensive?»
18	Взаимодействия между объектами (QT_Interaction)	«effects of Oxygen Plasma on stainless steel» «interaction between butyl rubber and glyceryl mono stearate»
19	Различия между объектами (QT_Difference)	«what is the difference between trisodium phosphate and sodium tripolyphosphate?» «tap drill depth versus thread depth»
20	Сходства между объектами (QT_Similarity)	«what are the similarities of dust filter and pumps?» «how similar is oxygen to hydrogen?»
21	Способы улучшения концепта (QT_Improvements)	«how to improve lubrication»
22	Методы реализации процесса, технологии (QT_Method)	«how to measure rotation?» «How to increase brake friction?» «methods for producing hydrogen»
23	Причины события, эффекта (QT_Cause)	«what are the causes of corrosion?» «reason of noise from cutter blade» «what can falsely trigger a sensor»
24	Предотвращение нежелательного события, эффекта (QT_Prevention)	«How to prevent reoxidation of steel» «prevention of restenosis» «how to avoid shin splints»
25	Условия протекания процесса (QT_Condition)	«copper melting conditions» «when can nitinol melt?»
26	Время действия (QT_Time)	«When was steam engine invented?»

27	Возможные сбои объекта (QT_Failure)	«what can go wrong with steam engine?» «failures of water pump»
28	Значения размера объекта (QT_Size)	«what is the size of dust particles?» «How big is a drillship?»
29	Значения длины объекта (QT_Length)	«how long is nanotube?» «length of aisle»
30	Значения высоты объекта или распространения действия (QT_Height)	«what is the height of a dental crown» «how high do unmanned vehicles fly»
31	Значения ширины объекта (QT_Width)	«width of strip» «how wide is casting bar?»
32	Значение расстояния между объектами (QT_Remoteness)	«how far is anode from cathode?» «distance between earphone speaker and ear»
33	Значение глубины объекта или действия (QT_Depth)	«how deep may a submarine dive?»
34	Значения продолжительности действия (QT_Duration)	«duration of heat pump drying»
35	Значения частоты объекта или действия (QT_Frequency)	«what is the frequency of X-ray?» «frequency of the electron beam» «how frequent is a shaft vibration?»
36	Значения массы объекта (QT_Mass)	«what is the mass of electron?»
37	Значения скорости объекта (QT_Speed)	«how fast is the shock wave?» «speed of rotary compressor»
38	Значения температуры объекта (QT_Temperature)	«what is the temperature of steam» «how hot is cold plasma?»
39	Значение толщины объекта (QT_Thickness)	«how thick is a wafer» «thickness of films»
40	Значение протяженности действия (QT_Distance)	«how far may airbus fly without stop?»
41	Значение возраста объекта (QT_Age)	«how old was the youngest cancer patient?»
42	Формы объекта (QT_Shape)	«shape of diesel nozzle hole?» «what is the shape of the pump impeller?»
43	Цвета объекта (QT_Color)	«what is a color of plasma?» «color of d-ribose»
44	Запахи объекта (QT_Smell)	«what is the smell of ball lightning?»

Следует отметить, что указанная классификация, несмотря на ее прямую ориентированность на предметные области, соответствующие инженерии знаний в инновационной деятельности, тем не менее включает наиболее востребованные в большинстве основных сфер жизнедеятельности типы атрибутивных знаний. Более того, использование в ней обобщенных классов, таких как QT_Object, QT_Subject, QT_Function, QT_Method и т. д., позволяет эффективно реализовать поиск по специфическим для других систем типам запросов, не включенных в данную классификацию. Так, например, запросы вида «What-River» и «What-Animal», определяющие в [6] различные типы ответов RIVER и ANIMAL соответственно, в данной классификации соответствуют общему типу QT_Instantiation, а также QT_Object/QT_Subject, с дополнительным ограничением по онтологическому классу ответа.

Полученная таким образом классификация может быть использована при разборе запросов пользователя с вычислением класса выраженной в них информацион-

ной потребности методом распознающих лингвистических шаблонов. Кроме того, указанные классы могут также использоваться при распознавании типов атрибутивных знаний, извлекаемых из текста на этапе индексации текстовых документов с применением семантического анализа, для их соотнесения с классом запроса, что составляет ключевую идею эффективного по производительности и качеству решения задачи информационного поиска в целом, в определенной степени моделирующего мыслительную деятельность человека при чтении текста. Данный подход, позволяющий преодолеть указанные выше недостатки существующих методов, был успешно применен при реализации вопросно-ответной системы Goldfire.

ЛИТЕРАТУРА

1. *Srihari, R.* A hybrid approach for named entity and sub-type tagging / Ro. Srihari, C. Niu, Wei Li // Proceedings of the 6th conference on Applied NLP. Stroudsburg, PA, USA, 2000. P. 247–254.

2. *Shen, D.* Our Winning Solution to Query Classification in KDDCUP 2005 / Dou Shen, Rong Pan, Jian-Tao Sun, Jeffrey Junfeng Pan, Kangheng Wu, Jie Yin, Qiang Yang // 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. SIGKDD Explorations, 2005. Vol. 7/2. P. 100–110.

3. *Lehnert, W. G.* A Conceptual Theory of Question Answering // Readings in natural Language Processing. Los Altos, CA, USA, Morgan Kaufmann Publishers Inc., 1986. P 651–658.

4. *Hermjakob, U.* Parsing and Question Classification for Question Answering / Proceedings of the workshop on Open-domain question answering. Stroudsburg, PA, USA, 2001. Vol. 12. P. 1–6.

5. *Bu, F.* Function-Based Question Classification for General QA / Fan Bu, Xingwei Zhu, Yu Hao, Xiaoyan Zhu / Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Cambridge, MA, USA, 2010. P. 1119–1128.

6. *Dubien, S.* Question Answering Using Document Tagging and Question Classification: A Thesis Master of Science // University of Lethbridge, Computer Science, 2003.