ТЕЗАУРУС ДЛЯ РАСШИРЕНИЯ ЗАПРОСОВ К МАШИНАМ ПОИСКА ИНТЕРНЕТА: СТРУКТУРА И ФУНКЦИИ[1]

П.И. Браславский, pb@imach.uran.ru
ИМаш УрО РАН

Мотивация

Исходным пунктом процесса поиска информации является информационная потребность, которая возникает у пользователя как следствие недостатка имеющейся информации для решения новой проблемы. Чтобы найти необходимую информацию, пользователь обращается к информационно-поисковой системе (ИПС). При обращении к ИПС пользователь должен сформулировать информационную потребность в виде запроса. Формулировка информационной потребности на языке запросов – трудно формализуемый этап поиска. Без опыта и знания принципов работы ИПС, при отсутствии представлений о коллекции документов сформулировать эффективный запрос довольно сложно.

Наиболее ярко эта ситуация проявляется при поиске в Интернете. В отличие от традиционной библиотеки, где можно обратиться за помощью к библиографу, при обращении к машинам поиска (МП) Интернета пользователь оказывается «один на один» с системой. Часто на формулировку запроса, который возвращал бы необходимый результат, уходит много времени. При этом развитые возможности языков запросов, как правило, остаются невостребованными, запросы к МП Интернета обычно состоят из двух-трех слов. Стандартом интерфейса МП Интернет *de facto* стало одно поле ввода, а наиболее распространенной формой запроса – несколько слов через пробел. Разработчики МП осознают бесперспективность «воспитания пользователя», поэтому в массе не стремятся к развитию языков запросов. Их усилия направлены на учет неявной информации, которая содержится в запросе, а также выявление предпочтений и ожиданий «массового пользователя».

Это не всегда идет на пользу «искателю» со специфическими информационными потребностями. Эти (относительно немногочисленные) пользователи формируют основное разнообразие запросов к МП. Если сообщество таких пользователей с близкими интересами не может позволить себе специализированный поисковый сервис, можно попытаться устранить дисбаланс между универсальностью МП и специфичностью информационных потребностей на этапе формулировки запросов.

Переформулировка (в частности – расширение) запросов – известный прием в практике информационного поиска. Один из вариантов расширения запросов – функция «найти похожие документы», представленная на многих МП. Эта функция позволяет дополнить первоначальный запрос словами указанного документа. Другой пример: некоторое время на МП AltaVista (www.altavista.com) был представлен сервис *AltaVista Refine*, который предоставлял пользователю возможность уточнить запрос с помощью словаря совместной встречаемости слов [Schwarz].

Традиционно в информационном поиске для модификации запросов использовались семантические словари – тезаурусы [Солтон]. Сегодня тезаурусы практически не находят применения в универсальных полнотекстовых МП Интернет. Одна из причин – в том, что чрезвычайно трудно построить тезаурус, который соответствовал бы тематическому разнообразию информации, индексируемой универсальной МП.

Наше предложение состоит в том, чтобы использовать тезаурус узкой предметной области в качестве основы независимой программы-ассистента формирования запросов к МП [Браславский, 2001]. Использование ассистента

[1] Работа поддержана грантом РФФИ 03-07-90342 и грантом конкурса научных проектов молодых ученых и аспирантов УрО РАН 2003 г.

позволяет сделать процесс формулировки запросов более осмысленным и эффективным. Такой ассистент может стать в свою очередь элементом предметно-ориентированной метапоисковой машины (МПМ) [Браславский, 2002].

В данной работе мы опишем наши подходы к разработке структуры тезауруса и описанию формата его представления.

Подход

Наш подход к описанию терминологии с помощью тезауруса во многом опирается на работы [Никитина, 1978, 1987, 1996]. Свойства терминов предметной области – системность, устойчивость и регулярность связей, отсутствие экспрессии, установка на объективность описания – делают возможным адекватное описание терминологии с помощью тезаурусов. Ключевой момент такого подхода – учет системных свойств терминов предметной области (понятийной структуры терминологии по [Шелов, 2001]). Еще один важный аспект подхода – использование в тезаурусе не только универсальных (например, «род-вид», «часть-целое» и т.д.), но и специфических для конкретной предметной области отношений. Таким образом, не только термины тезауруса (узлы), но и связи несут значительную семантическую нагрузку.

Основным элементом тезауруса мы считаем *концепцию*, т.е. понятие, которое выражается термином, а не сам термин (в этом наш подход отличается от подхода С.Е. Никитиной). Такое решение позволяет естественным образом снять проблему описания полисемии: значение термина определяется концепцией, которой он соответствует. Кроме того, отпадает необходимость дифференциации и описания различных типов эквивалентности терминов (синонимии, частичной синонимии, иностранных эквивалентов): одной концепции может соответствовать несколько терминов (в т.ч. разноязычных). Подход на основе концепций позволяет эффективно управлять гранулярностью («зернистостью») описания: разработчик тезауруса самостоятельно может определить необходимый порог семантического сходства/различий терминов за счет укрупнения/сужения концепций. Нам представляется, что такой подход хорошо соответствует нашей цели – разработке структуры тезауруса для модификации информационных запросов, – хотя может оказаться слишком грубым или упрощенным для других лексикографических задач.

Мы должны предоставить разработчику тезауруса свободу и в другом вопросе – определении набора отношений тезауруса. Семантические отношения должны максимально соответствовать предметной области тезауруса. Дифференциация возможных типов семантических отношений особенно важна для разработки полуавтоматических процедур модификации запросов. В пределе, если в тезаурусе представлен только один тип связи – ассоциация, – поле для применения автоматизированных методов сильно сокращается.

При разработке мы ориентировались на то, что тезаурус и соответствующий формат представления должны одновременно выполнять несколько функций.

Репрезентативная функция – адекватно (с точки зрения поставленной прикладной задачи) описывать терминологию предметной области. На этом уровне ключевым моментом является выделение структурных элементов описания. Примером формата, соответствующего этой функции, является DSL (Dictionary Specification Language), используемый в словарной системе Lingvo компании ABBYY Software House. DSL является по преимуществу средством описания внешнего вида (репрезентации) словарной статьи [Ассоциация]. На этом же уровне можно рассматривать бумажные словари.

Прикладная функция. Тезаурус при нашем подходе является составной частью программы-ассистента. Важная задача – сделать процедуры работы с тезаурусом независимыми от конкретного тезауруса. Для этого структура тезауруса должна обладать внутренней интерпретируемостью.

Формат должен выполнять также *коммуникативную функцию*, т.е. способствовать повторному использованию, обмену и интеграции терминологических данных в виде тезаурусов.

Кроме того, при разработке перед нами стояла традиционная задача сохранения баланса между развитостью структуры, выразительными возможностями формата – с одной стороны – и простотой, прозрачностью описания – с другой.

Исходя из этих требований, в качестве формата представления тезауруса мы выбрали язык XML. Формат тезауруса описывается в виде XML Schema [XML].

В качестве аналогов при работе мы рассматривали формат словаря Virtual HyperGlossary и документы проекта SALT [Virtual, SALT]. При разработке формата описания тезауруса мы старались ориентироваться на отечественный и международный стандарты [ГОСТ, ISO]. В частности, мы стремились привести словарь описания тезауруса к словарю стандарта ISO 12620: 1999 Computer applications in terminology – Data categories.

Структура тезауруса

В этом разделе мы кратко опишем структуру тезауруса на основании диаграмм XML Schema. Саму схему, а также более детальное ее описание и примеры можно найти по адресу http://imach.uran.ru/pb/thesaurus/.

Тезаурус представляет собой совокупность концепций – элементов *conceptEntry* (рис. 1). Заголовочная часть тезауруса состоит из обязательных элементов: описания тезауруса (*thesaurusDescription*), даты (*data*), информации о создателях (*originatingEntity*), кода УДК (*UDC*). Необязательные элементы заголовочной части – библиографические ссылки на источники, использованные при составлении тезауруса (*sourceIdentifier*), и комментарии (*comment*)

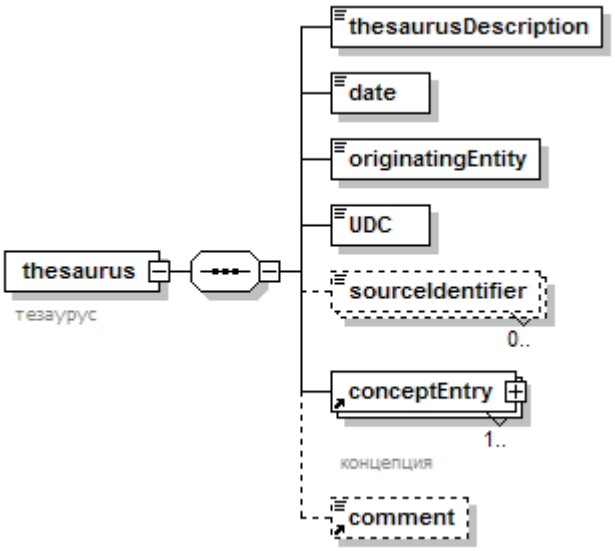

Рис 1. Структура коневого элемента тезауруса

Концепции (*conceptEntry*) соответствуют терминологические описания (*termEntry*), а также характеристика (*characteristic*), категория (*category*), определение (*definition*), набор связей (*link*), ссылка на нетекстовый иллюстративный материал (*figure*) и комментарии (*comment*) (рис. 2). Элемент *characteristic* заимствован из [ISO]. Характеристика служит для формирования и уточнения концепции, ее роль аналогична роли релятора и лексического примечания [ГОСТ]. Элемент *category* указывает на категориальную принадлежность концепции ([ISO] не предусматривает категорию данных с таким именем). Значение *category* используется для построения систематического указателя и принадлежит типу *categoryType*. Отечественный стандарт [ГОСТ] предлагает использовать следующие общие категории:

- названия дисциплин и отраслей деятельности (*subject*);
- предметы, материалы (object);
- методы, процессы, операции, явления (*process*);
- свойства, величины, параметры, характеристики (*property*);
- отношения, структуры, модели, законы, правила, абстрактные понятия (*abstract*).

Среди атрибутов элемента *conceptEntry* есть обязательный атрибут *conceptEntryId*, который служит уникальным идентификатором концепции. Элемент *link* – «пустой», но содержит два атрибута: факультативный – имя связи (*name*) и обязательный – идентификатор связанной концепции (*conceptEntry*). Атрибут *name* принадлежит типу *linkType*. Все элементы *conceptEntry*, кроме *termEntry*, являются необязательными.

Типы *categoryType* и *linkType* являются производными от строчного типа (*string*). Оба типа заданы перечислением допустимых значений. На данный момент тип *categoryType* задан перечислением пяти категорий, соответствующих [ГОСТ] (см. выше). Тип *linkType* включает только три базовых значения – «род» (*broaderConceptGeneric*), «целое» (*broaderConceptPartitive*) и «ассоциация» (*relatedConcept*).

Схема устанавливает уникальность идентификаторов концепций (*conceptEntryId*) и связность ссылок *link* (с помощью механизмов *key* и *keyref*), а также уникальность (*unique*) идентификаторов концепций в наборе *link* (нельзя установить более одной связи между двумя концепциями).

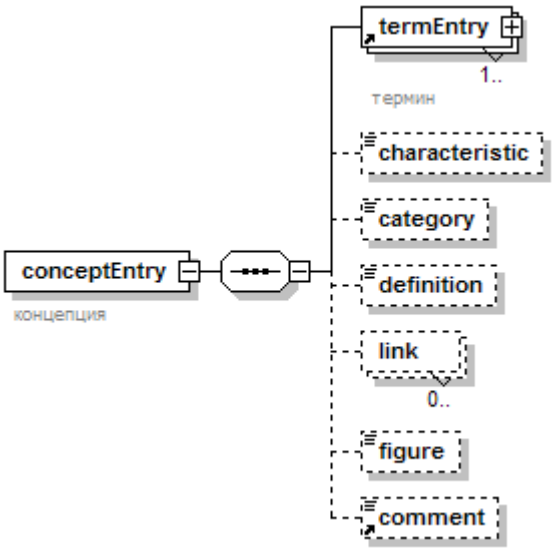

Рис 2. Структура элемента «Концепция»

Наконец, элемент *termEntry* имеет обязательный атрибут «язык» (*lang*) и включает в себя обязательный элемент «термин» (*term*). Грамматическая информация может содержаться в атрибуте *gramInfo* элемента *term*. Нам не удалось найти единообразного способа описания грамматической информации, принятого отечественными разработчиками электронных словарей, поэтому тип этого атрибута – строчный (*string*). Терминологическое описание могут дополнять сокращения (*acronym*), варианты (*variant*), однокоренные слова (*cognate*), примеры использования (*context*), числовая характеристика (*weight*), а также комментарии (*comment*). Семантика числовой характеристики *weight* пока не зафиксирована. Это может быть вес, экспертная оценка важности для задач поиска или статистический параметр (например, *idf* по какой-нибудь коллекции). Отметим, что все терминологические описания, соответствующие одной концепции, равноправны.

Как мы уже сказали выше, мы предлагаем делегировать разработчикам тезауруса не только функции наполнения тезауруса, но и определения набора допустимых семантических связей между концепциями. Такой набор должен наилучшим образом подходить терминологии описываемой предметной области. Возможные значения соответствующего атрибута типа *linkType* элемента *link* задаются с помощью перечисления. Мы предполагаем, что разработчики при необходимости будут использовать механизм *redefine* переопределения типа в рамках пространства имен (*namespace*), задаваемого основной схемой.

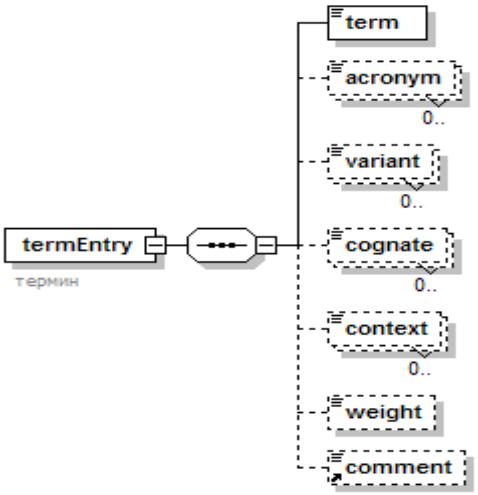

Рис 3. Структура элемента «Термин»

```
<conceptEntry conceptId="1234">
    <termEntry lang="ru">
        <term>автоматический оптический контроль</term>
        <acronym>АОК</acronym>
        <variant>автоматизированный оптический контроль</variant>
        <context>Система АОК состоит из оптико-электронного блока и системы анализа видеоинформации.</context>
    </termEntry>
    <termEntry lang="en-us">
        <term>automatic optical inspection</term>
        <acronym>AOI</acronym>
        <context>Automatic optical inspection (AOI) systems detect the same type of surface-related defects as manual inspection, including bare-board inspection, solder bridging, lack of solder, missing components, poor part orientation, lifted leads, tombstoning, and solder balls.</context>
    </termEntry>
    <category>process</category>
    <definition>Разновидность неразрушающего контроля, который использует методы обработки изображений, распознавания образов, машинного зрения и др., чтобы на основе изображения объекта установить его соответствие технологическим допускам.</definition>
    <link name="relatedConcept" conceptEntry="5678"/>
    <link name="broaderConceptGeneric" conceptEntry="5690"/>
    <!--концепция связана с двумя другими-->
</conceptEntry>
```

Рис. 4. Фрагмент описания тезауруса

В настоящее время в Институте машиноведения УрО РАН ведется разработка тезауруса предметной области «Оптический контроль изделий микроэлектроники». В качестве примера на рис. 4 приведено описание одной из концепций верхнего уровня – «автоматический оптический контроль».

Заключение

Разработка формата представления терминологии для задачи модификации запросов к машинам поиска Интернет еще не завершена. Однако нам представляется, что предложенный подход позволяет получить одновременно лаконичное и выразительное описание тезауруса, соответствующее поставленным требованиями. Дальнейшее развитие формата нам видится в уточнении, структуризации и стандартизации отдельных элементов описания. Это касается, например, библиографических ссылок (*sourceIdentifier)* и описания грамматических характеристик термина (*gramInfo*). Кроме того, в будущем необходимо предусмотреть возможность интеграции тезаурусов, в частности – возможность установления связей между концепциями разных тезаурусов.

Прикладные задачи повышения эффективности поиска информации в Интернет заставили вторгнуться в новую для нас область – область терминоведения и лексикографии. Мы надеемся, что замечания, вопросы, предложения и рекомендации, касающиеся затронутого круга вопросов, сделают нашу работу более результативной.

Литература

1. Ассоциация лексикографов Lingvo – http://www.lingvoda.ru/
2. Браславский П.И. Метапоисковая машина для поиска специализированной научной информации в интернете: структура и функции // Вестник Томского гос. ун-та, – 2002. – №1 (II) Приложение. – С. 353-356.

3. Браславский П.И. Построение запросов к машине поиска Internet с помощью тезауруса // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Сб. докладов Третьей Всероссийской конференции RCDL'2001. – Петрозаводск: КарНЦ РАН, 2001. – С. 83-87.
4. ГОСТ 7.25 – 2001. Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления. – М.: Изд-во стандартов, 2001.
5. Никитина С.Е. Васильева Н.В. Экспериментальный системный словарь стилистических терминов. Принципы составления и избранные словарные статьи. – М., 1996. – 172 с.
6. Никитина С.Е. Семантический анализ языка науки. (На материале лингвистики.) – М.: Наука, 1987. – 276 с.
7. Никитина С.Е. Тезаурус по теоретической и прикладной лингвистике. – М.: Наука, 1978. – 312 с.
8. Солтон Дж. Динамические библиотечно-информационные системы. – Пер. с англ. – М.: Мир, 1979. – 558 с.
9. Шелов С.Д. Терминоведение: семь вопросов и семь ответов по семантике термина // НТИ. Сер. 2. Информ. процессы и системы, – 2001. – №2. – С. 1-12.
10. ISO 12620: 1999 Computer applications in terminology – Data categories.
11. SALT project — XML representations of Lexicons and Terminologies (XLT) — Default XLT Format (DXLT) – http://www.ttt.org/oscar/xlt/dxltspecs.html
12. Schwarz C. Web Search Engines // Journal of the American Society for Information Science, – 1998. – №49 (11). – P. 973-982.
13. Virtual HyperGlossary – http://www.vhg.org.uk/home/
14. XML Schema – http://www.w3.org/XML/Schema