# Knowledge Engineering

## К ВОПРОСУ ЕСТЕСТВЕННО-ЯЗЫКОВОЙ АДРЕСАЦИИ

**Крассимира Иванова, Виталий Величко, Крассимир Марков**

***Аннотация**: В настоящей работе представлена идея естественно-языковой адресации. Это дополнительная возможность для представления онтологической информации в интеллектуальных системах. Естественно-языковая адресация имеет ряд преимуществ. На первом месте – это линейная алгоритмическая сложность, которая зависит от максимальной длины слов (max_L), а не от их количества. Во-вторых, это уменьшение объема занимаемой памяти – дополнительные индексы не используются. В-третьих, уменьшение времени обработки из-за полного отсутствия поиска – информация извлекается прямо по адресу. Необходимо отметить, что это универсальное представление информации одновременно доступной как для человека, так и для автоматизированных систем. Такой способ организации информации применим для ее хранения и использования в библиотеках онтологий, терминов, понятий, текстовых документов.*

***Ключевые слова**: Естественно-языковая адресация, организация онтологических баз данных.*

***ACM Classification Keywords**: D.4.2 Storage Management; E.2 Data Storage Representations*

### Введение

Растущее развитие средств для оперирования с онтологиями и базами знаний определяет устойчивую тенденцию увеличения объема мета-информации, создаваемой людьми и компьютерами. С момента появления Semantic Web [sw, 2012] практически любой объект в Интернете содержит значительное количество метаданных, описывающих различные семантические аспекты. Наличие стандартизированных языков для описания метаданных (XML, SHOE, DAML+OIL, RDF и RDFS и др.) делает мета-информацию доступной для анализа и интерпретации.

Увеличение количества метаданных является феноменом, который требует особого внимания и анализа. Бытует иллюзорное впечатление, что в глобальной сети уже есть все, что нас интересует, но пока не описано полностью мета-данными и является вопросом времени, чтобы это произошло. Когда это случиться, семантико-поисковые системы будут находить для нас необходимую информацию, которая должна быть и “достаточной”.

Более чем очевидно, что "Никто не обнимет необъятного" [Prutkov, 1863], особенно если к нему добавлено "мета-необъятное". Поэтому поиск эффективных решений для работы с естественными структурами языка и их смыслом остается серьезной проблемой. Наше внимание в этой работе направлено на поиск новых возможностей хранения онтологических структур, основанных на специфической "естественно-языковой адресации".

## Пример адресации в системе WordNet

**WordNet** — это семантическая сеть для английского языка, разработанная в Принстонском университете, и свободно доступная вместе с сопутствующим программным обеспечением [WordNet, 2012]. WordNet можно свободно использовать в коммерческих и научных целях. Для работы с ним существует несколько программ, множество интерфейсов и API, реализованное как на большинстве возможных языков, так и с помощью протокола DICT, программы GoldenDict и других. Пакеты WordNet присутствуют в некоторых репозиториях GNU ПО, Linux и их дистрибутивах.

Рассмотрим организацию информации в системе WordNet. Для примера мы выбрали слово „accession".

На запрос о слове „accession", система WordNet выдает следующую информацию (рис. 1):

The noun accession has 6 senses (no senses from tagged texts)

1. {13251723} <noun.process> accession#1 -- (a process of increasing by addition (as to a collection or group); "the art collection grew through accession")
2. {13170404} <noun.possession> accession1#2 -- ((civil law) the right to all of that which your property produces whether by growth or improvement)
3. {13082910} <noun.possession> accession#3, addition#4 -- (something added to what you already have; "the librarian shelved the new accessions"; "he was a new addition to the staff")
4. {07078650} <noun.communication> accession2#4, assenting#1 -- (agreeing with or consenting to (often unwillingly); "accession to such demands would set a dangerous precedent"; "assenting to the Congressional determination")
5. {05115154} <noun.attribute> entree#2, access#1, accession#5, admittance#1 -- (the right to enter)
6. {00232781} <noun.act> accession3#6, rise to power#1 -- (the act of attaining or gaining access to a new office or right or position (especially the throne); "Elizabeth's accession in 1558")

The verb accession has 1 sense (no senses from tagged texts)

1. {00989696} <verb.communication> accession#1 -- (make a record of additions to a collection, such as a library)

Рис. 1 Ответ системы WordNet на запрос о слове „accession"

Рассмотрим, как именно получен этот ответ в системе WordNet. Словарь WordNet состоит из 4 сетей для основных знаменательных частей речи: существительных, глаголов, прилагательных и наречий. Базовой словарной единицей в WordNet является не отдельное слово, а так называемый синонимический ряд («синсеты»), объединяющие слова со схожим значением и по сути своей являющимися узлами сети. Для удобства использования словаря человеком каждый синсет дополнен определением и примерами употребления слов в контексте. Слово или словосочетание может появляться более чем в одном синсете и иметь более одной категории части речи. Каждый синсет содержит список синонимов или синонимичных словосочетаний и указатели, описывающие отношения между ним и другими синсетами. Слова, имеющие

несколько значений, включаются в несколько синсетов и могут быть причислены к различным синтаксическим и лексическим классам.

WordNet хранит информацию в четырех основных файлах данных (существительные, глаголы, прилагательные и наречия). В каждом из этих файлов структура данных одинакова – для каждого слова хранится один или несколько наборов синонимических множеств (синсетов), доступ к которым осуществляется по адресу первого байта синсета, который задан явно с помощью восьми десятичных знаков записанных с первого байта синсета (рис. 2 и рис. 3). Данные в синсете разделены пробелами. После адреса синсета следуют три служебных полей, а вслед за ними само слово, после него следует другая лингвистическая информация. Необходимо отметить, что ссылки на другие синсеты задаются снова через их абсолютные адреса в файле. (Элементы, которые нас интересуют, выделены жирным шрифтом).

**13251723** 22 n 01 **accession** 0 001 @ **13323403** n 0000 | a process of increasing by addition (as to a collection or group); "the art collection grew through accession"

**13170404** 21 n 01 **accession** 1 002 @ **13070995** n 0000 ;c **08338303** n 0000 | (civil law) the right to all of that which your property produces whether by growth or improvement

**13082910** 21 n 02 **accession** 0 addition 0 001 @ **13082742** n 0000 | something added to what you already have; "the librarian shelved the new accessions"; "he was a new addition to the staff"

**07078650** 10 n 02 **accession** 2 assenting 0 002 @ **07076600** n 0000 + **00795631** v 0102 | agreeing with or consenting to (often unwillingly); "accession to such demands would set a dangerous precedent"; "assenting to the Congressional determination"

**05115154** 07 n 04 entree 0 access 0 **accession** 0 admittance 0 003 @ **05113619** n 0000 + **02426186** v 0401 ~ **05119817** n 0000 | the right to enter

**00232781** 04 n 02 **accession** 3 rise_to_power 0 003 @ **00060914** n 0000 + **01989112** v 0101 + **02358456** v 0101 | the act of attaining or gaining access to a new office or right or position (especially the throne); "Elizabeth's accession in 1558"

Рис. 2. Синсеты слова „accession” в WordNet файле с данными для существительных

**00989696** 32 v 01 **accession** 0 002 @ **00990286** v 0000 ;c **00897092** n 0000 01 + 08 00 | make a record of additions to a collection, such as a library

Рис. 3. Синсеты слова „accession” в WordNet файле данных для глаголов

Для нас важно, как выполняется доступ к данному синсету. Очевидно, что необходимо где-то хранить информацию о том, где находятся синсеты для каждого слова. Это делается через WordNet индексные файлы, которых тоже четыре, в соответствии с файлами данных (существительные, глаголы, прилагательные и наречия). Они отсортированы в алфавитном порядке слов, каждому слову соответствует одна строка в индексном файле, содержащая слово, краткую служебную информацию и абсолютные адреса всех синсетов, в которые входит слово (рис. 4 и рис. 5).

```
accession n 6 4 @ ~ + ; 6 0 13251723 13170404 13082910 07078650
05115154 00232781
```

Рис. 4. Строка для слова „accession” в WordNet индексном файле с данными для существительных

```
accession v 1 2 @ ; 1 0 00989696
```

Рис. 5. Строка для слова „accession” в WordNet индексном файле с данными для глаголов

Чтобы прочитать все синсеты для данного слова, сначала выполняется бинарный поиск во всех индексных файлах, а затем через прямой доступ по абсолютным адресам читаются данные из файлов. Алгоритмическая сложность в данном случае $O(n_n*lg(n_n)+n_v*lg(n_v)+n_a*lg(n_a)+n_r*lg(n_r))$, где $n_n$, $n_v$, $n_a$ и $n_r$ представляют собой количества существительных, глаголов, прилагательных и наречий.

Существует и второй способ получения абсолютных адресов синсетов. Он выполняется с помощью т.н. индекса смыслов (sense index). Этот индекс тоже отсортирован, но каждое слово записывается в таком количестве строк, сколько синсетов существуют для этого слова во всех файлах данных. Например, слово „accession” имеет семь строк в индексе смыслов - шесть для его значений как существительное и одна - в качестве глагола. Каждая строка содержит только один абсолютный адрес синсета в соответствующем файле данных (рис. 6).

```
accession%1:04:03:: 00232781 6 0
accession%1:07:00:: 05115154 5 0
accession%1:10:02:: 07078650 4 0
accession%1:21:00:: 13082910 3 0
accession%1:21:01:: 13170404 2 0
accession%1:22:00:: 13251723 1 0
accession%2:32:00:: 00989696 1 0
```

Рис. 6 Строки слова „accession” в индексе смыслов

Для получения всех синсетов данного слова в индексе смыслов, выполняется сначала бинарный поиск, затем просматриваются все имеющиеся для данного слова строки, и, наконец, по всем полученным адресам непосредственно считывают синсеты из файлов данных. Алгоритмическая сложность в данном случае больше, чем $O(n*lg(n))$, где $n = n_n+n_v+n_a+n_r$, т.е. n - общее количество слов в базе данных (существительные+глаголы+прилагательные+наречия), так как из-за множества значений, слова могут повторяться много раз, что приводит к увеличению количества операций по нахождению всех строк этого слова.

## Естественно-языковая адресация

Организация информации в системе WordNet позволяет быстро получить необходимую информацию, используя „одновременный” бинарный поиск в четырех стандартных индексах или в одном индексе смыслов, а затем выполнить позиционирование непосредственно в соответствующих файлах данных.

В данном случае необходимо отметить некоторые недостатки такой организации:

1) абсолютная адресация удобна для компьютерной обработки, но не удобна для пользователя;

2) ручное формирование абсолютных адресов невозможно, а их использование не может быть реализовано без поддержки соответствующей программы;

3) пользователю предоставляется статическая ("скомпилированная") версия базы данных, которую невозможно развивать - она пригодна только для чтения.

Любые изменения, приводящие к изменению количества байтов в основном файле данных, делает его непригодным для использования из-за абсолютных адресов, являющимися указателями. Например, на рис. 7 показан синсет слова „accession" из (а) настоящей и (б) более ранней версии WordNet файлов данных для существительных. Более ранняя версия синсета опубликована в [Palagin et al, 2011]. На рис. 7 (а) и (б) видна разница в адресах.

```
13082910 21 n 02 accession 0 addition 0 001 @ 13082742 n 0000 |
     something added to what you already have; "the librarian
     shelved the new accessions"; "he was a new addition to the
     staff"
                                a)

00047131 04 n 02 accession 0 addition 0 001 @ 09536731 n 0000 |
     something added to what you have already; "the librarian
     shelved the new accessions"; "he was a new addition to the
     staff"
                                б)
```

Рис. 7 Синсет слова „accession" из (а) настоящей и (б) более ранней версии файла для существительных

Это означает, что любое изменение в информации, даже на один байт, требует полной рекомпиляции соответствующей части базы данных – основного файла, его индекса, индекса смыслов, который соответствует всем основным файлам.

Если проанализируем структуру синсета, увидим одну очень важную особенность.

В памяти компьютера все символы представляются цифровыми кодами (занимающими один, два или четыре байта, в зависимости от системы кодирования - ASCII или Unicode). Таким образом, один уникальный цифровой код (абсолютный адрес) указывает на другой, тоже уникальный цифровой код (компьютерное представление слова, например в ASCII-кодировке слово „accession" имеет следующее представление: 97 99 99 101 115 115 105 111 110).

Этого можно избежать, если использовать другой тип организации информации.

Внутреннее представление (код) слова может быть непосредственно использовано при построении онтологии. Этот код можно рассматривать как пространственный адрес (в девятимерном пространстве в случае со словом „accession"). По адресу слова можно сохранить всю информацию синсета и получить ее снова через этот адрес.

Для людей адрес слова „accession" будет представлен самим словом „accession", а для компьютера через вектор (97, 99, 99, 101, 115, 115, 105, 111, 110).

Этот способ адресации назовем „***естественно-языковая адресация***".

Учитывая, что в естественном языке слова имеют разную длину, а некоторые словосочетания являются понятиями, возникает требование возможности одновременной работы с взаимосвязанными

информационными пространствами различной размерности. Такую возможность предоставляет „Мультидоменная информационная модель" [Markov, 2004] и соответствующее программное обеспечение, названное - „ Мультидоменный метод доступа" [Markov, 1984].

Если применить эту возможность к WordNet базе данных, получим результат более понятный для человека (Рис.8. б) и в тоже самое время – полностью понимаемый компьютером.

```
13082910 21 n 02 accession 0 addition 0 001 @ 13082742 n 0000 |
     something added to what you already have; "the librarian
     shelved the new accessions"; "he was a new addition to the
     staff"
                                a)

accession 21 n 02 ; 0 addition 0 001 @ acquisition n 0000 |
     something added to what you already have; "the librarian
     shelved the new accessions"; "he was a new addition to the
     staff"
                                б)
```

Рис. 8 Синсет слова „accession" из (а) настоящей и (б) NL-версии WordNet файла для существительных

## Заключение

В настоящей работе была представлена идея естественно-языковой адресации. Это дополнительная возможность для представления онтологической информации в интеллектуальных системах. Она имеет ряд преимуществ. На первом месте это линейная алгоритмическая сложность, которая зависит от максимальной длины слов (max_L), а не от их количества, т.е. O(max_L). Во-вторых, это уменьшение объема занимаемой памяти из-за полного отсутствия дополнительных индексов, абсолютных адресов и дополнительных файлов. Можно указать на уменьшение времени обработки вследствие полного отсутствия поиска – информация извлекается по прямому адресу. И не на последнем месте – универсальное представление информации одновременно доступной как для человека, так и для автоматизированных систем.

Этот способ организации информации используется в „Инструментальном комплексе онтологического назначения" (ИКОН) [Palagin et al, 2011], который разрабатывается в Институте кибернетики им. В.М.Глушкова НАНУ, Киев. Естественно-языковая адресация предусмотрена для хранения и использования информации в библиотеках онтологий, терминов и понятий, текстовых документов.

## Благодарности

Работа опубликована при финансовой поддержке проекта **ITHEA XXI** Института информационных теорий и приложений FOI ITHEA Болгария www.ithea.org и Ассоциации создателей и пользователей интеллектуальных систем ADUIS Украина www.aduis.com.ua.

## Библиография

[Markov, 1984] Kr.Markov. A Multi-domain Access Method. // Proceedings of the International Conference on Computer Based Scientific Research. Plovdiv, 1984. pp. 558-563.

[Markov, 2004] Markov, K. Multi-domain information model. Int. J. Information Theories and Applications, 11/4, 2004, pp.303-308.

[Palagin et al, 2011] А.В. Палагин, С.Л. Крывый, Н.Г. Петренко. Онтологические методы и средства обработки предметных знаний: монография/Луганск: изд-во ВНУ им. В. Даля, 2011. – 323 с.

[Prutkov, 1863] Прутков Козьма Петрович: Сочинения. 1863г. [Электронный ресурс] // URL: http://az.lib.ru/p/prutkow_k_p/ (дата обращения 04.08.2012).

[sw, 2012] Semantic Web: [Электронный ресурс] // URL: http://www.w3.org/2001/sw/ (дата обращения 04.08.2012).

[WordNet, 2012] WordNet®. A lexical database for English. Princeton University. [Электронный ресурс] // URL: http://wordnet.princeton.edu/ (дата обращения 04.08.2012).

## Информация об авторах

***Krassimira Ivanova*** *– University of National and World Economy, Sofia, Bulgaria*
*e-mail: krasy78@mail.bg*
*Major Fields of Scientific Research: Data Mining*

***Vitalii Velychko*** *– Institute of Cybernetics, NASU, Kiev, Ukraine*
*e-mail: velychko@aduis.com.ua*
*Major Fields of Scientific Research: Data Mining, Natural Language Processing*

***Krassimir Markov*** *– Institute of Mathematics and Informatics at BAS, Sofia, Bulgaria;*
*e-mail: markov@foibg.com*
*Major Fields of Scientific Research: Multi-dimensional information systems, Data Mining*