УДК 004.89:004.823

МЕТОД ПОСТРОЕНИЯ СЛОВАРЕЙ ПРЕДМЕТНЫХ ОБЛАСТЕЙ ДЛЯ ИЗВЛЕЧЕНИЯ ФАКТОВ ИЗ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

А.Б. Кунгурцев
Кандидат технических наук, профессор*
Контактный тел.: (0482) 68-03-02
E-mail: abkun@te.net.ua

С.Н. Бородавкин
Аспирант*
Контактный тел.: (0482) 44-38-59
E-mail: lw0000@inbox.ru

А.П. Голуб
Магистр*
Контактный тел.: 066-606-17-88
E-mail: bellerafont@gmail.com
*Кафедра системного программного обеспечения
Институт компьютерных систем Одесского
национального политехнического университета
пр-т Шевченко, 1, Одесса, Украина, 65044

Розроблено метод побудови словника термінів, що базується на попередній обробці множини текстів з заданої предметної області. Описана побудова множини фреймів, що визначають модель знань предметої області, на основі отриманого словника

Ключові слова: словник, предметна область, фрейм, модель знань

Разработан метод построения словаря терминов, основанный на предварительной обработке множества текстов из заданной предметной области. Описано построение множества фреймов, определяющих модель знаний предметной области, на основе полученного словаря

Ключевые слова: словарь, предметная область, фрейм, модель знаний

A method to build a dictionary of terms is developed, which is based on a processing of preliminary set of texts from the given domain. Building a set of frames is described, which determine the domain knowledge model, based on a resulting dictionary

Key words: dictionary, domain, frame, knowledge model

1. Введение

В работе [1] выделены основные этапы процесса анализа текстов на естественном языке с целью выделения фактов:

1. Синтаксический разбор
2. Семантический анализ

Результатом синтаксического разбора является дерево разбора текста. При выполнении семантического анализа, пользуясь полученным деревом, необходимо построить модель знаний входного текста, представляющую собой сеть фреймов. В общем случае, выполнение такого построения является нетривиальной задачей, поскольку каждый фрейм должен отражать некоторое понятие из предметной области, а выделение таких понятий и соответствущих им терминов (слов, именных групп) сводится к:

- выделению имен существительных и отсечению служебных частей речи (предлогов, союзов и пр.), глаголов, прилагательных и т.д.
- составлению частотного словаря входного текста с целью выделению наиболее часто встречающихся понятий
- выполнению лингвистического анализа текста с учетом того, какой частью речи является рассматриваемое слово. Кроме того, необходимо произвести анализ взаимосвязей между словами в предложении, что позволит выделить устойчивые выражения.

Таким образом, для построения сети фреймов на основе произвольного входного текста, актуальной

является задача предварительного составления словаря предметной области, к которой относится анализируемый текст. Такие словари находят широкое применение в разработке информационных систем, создании объектного представления баз данных, систем поиска, а также справочных систем. Каждый из терминов такого словаря, фактически, представляет собой определение фрейма, а весь словарь – множество фреймов предметной области F_d. В дальнейшем, при выполнении семантического анализа текста из заданной предметной области, фреймы из словаря возможно наполнять содержимым (слотами). Очевидно, что модель знаний анализируемого текста может определять не все фреймы из F_d. Таким образом, после анализа текста, из F_d можно исключить фреймы, не получившие наполнения, получив множество F_t фреймов, раскрытых в тексте ($F_t \subseteq F_d$).

2. Подбор текстов для анализа

Создание словаря предметной области основывается на анализе текстов, которые содержат термины, специфические для заданной предметной области. Соответственно для создания качественного словаря предметной области, необходимо подобрать такой набор текстов, который бы содержал максимальное количество специфических для данной предметной области терминов и при этом содержал небольшое количество терминов из другой предметной области. Однако, в большинстве случаев тексты содержат большое количество служебных слов, терминов из других или смежных предметных областей.

Для формирования набора исходных текстов можно использовать два подхода:

- автоматизированный
- ручной(привлечение эксперта)

Определение исходного набора текстов автоматизированным путем, является практически неразрешимой задачей, так как для отбора текстов Тот из исходного множества Тисх необходимо определить границы предметной области и разработать критерии принадлежности к ней. Для решения данной задачи необходимо применять глубокий семантический анализ, максимально разрешать все неоднозначности при синтаксическом и семантическом разборах, снимать омонимию. Т.е., по сути, для автоматизированного определения набора текстов, необходимо реализовать алгоритмы, во много раз более сложные, чем применяемые при непосредственном построении словаря предметной области, и потому выходящие за рамки решаемой задачи.

Под ручным методом формирования исходного набора текстов подразумевается привлечение стороннего эксперта (человека или группу людей, имеющих обширные познания в анализируемой предметной области, а также смежных с ней).

При отборе текстов для составления словаря предметной области необходимо проанализировать саму структуру области. Например, если предметная область некоторой организации достаточно велика, то в ней наверняка присутствуют структурные подразделения – бухгалтерия, отдел кадров и т.д. Очевидно, что данные подразделения имеют свою специальную документацию и, соответственно, специфические термины предметной области. Таким образом, исходную предметную область можно разделить на несколько подобластей, каждая из которых характерна для определенного структурного подразделения организации.

Возможны ситуации, когда анализируемый текст имеет тезисное представление. Текст такого типа может содержать большое количество терминов предметной области, однако при этом частотные характеристики этих терминов будут невысоки, так как тезисы не допускают повторений и представляют информацию в максимально лаконичной форме. При анализе такого текста наряду с нетезисным текстом, возникает ситуация потери терминов с низкой частотой. Решить данную проблему можно введением весового коэффициента K_t, показывающего важность текста с точки зрения формирования словаря предметной области. Весовой коэффициент задается экспертом на стадии отбора текстов для анализа и представляет собой число в диапазоне [0;10]. Данный коэффициент используется как множитель при расчете частоты вхождения термина (1):

$$f_{сл} = \frac{N_{сл}}{N_{общ}} \cdot n_f \cdot K_t \qquad (1)$$

где $f_{сл}$ – частота вхождения, $N_{сл}$ – количество вхождений слова в анализируемый текст, $N_{общ}$ – общее количество слов в тексте, n_f - количество слов, по которому определяется частота вхождения, K_t – весовой коэффициент текста.

Выставляя повышенный коэффициент текстам, точно соответствующим предметной области, и уменьшенный, тем, что затрагивают смежные области, можно избежать попадания в окончательный словарь именных групп и слов, не относящихся к предметной области. Отсутствие процедуры предварительного подбора и оценки текстов приводило к существенному понижению качества словаря [2].

3. Выделение терминов

Термин – это слово, устойчивое словосочетание или сокращение, которое выражает и в известной мере классифицирует в данной предметной области определённое понятие или сущность, отражая в своей смысловой структуре характеристические признаки объекта терминирования и взаимосвязи этого объекта с другими с достаточной для взаимного общения точностью.

Термины, в отличие от общеупотребительных слов:

- выражают специальные понятия, имеют повышенную смысловую точность
- обладают свойством систематичности (отражают взаимосвязь понятий, появляющихся в процессе развития предметной области); пригодны к дальнейшему терминообразованию, т.е. к образованию производных терминов и употреблению в терминологических словосочетаниях;
- характеризуются краткостью.
- термин должен не только называть предмет, но и сообщать о нём, т.е. выполнять коммуникативную функцию языка.

Для выделения терминов из анализируемого текста можно использовать статистические и лингвистические методы анализа.

Статистическая модель выделения терминов основывается на расчете частотных характеристик анализируемого текста, т.е. по сути, составление частотного словаря. Главным недостатком частотных словарей, с точки зрения выделения терминов, является то, что они не учитывают, какой частью речи является рассматриваемое слово. В соответствии с [3], наиболее часто встречающимися частями речи являются союзы, предлоги, местоимения и т.п., которые, очевидно, не могут быть терминами. Это обусловлено тем, что со статистической точки зрения язык представляет собой большое количество редких событий [4], в результате чего небольшое количество слов встречается очень часто, а подавляющее большинство слов имеют очень невысокую частоту.

Таким образом, статистический анализ, для выделения терминов непригоден, поскольку необходимо производить анализ текста с учетом того, какой частью речи является рассматриваемое слово, а также выполнять анализ взаимосвязей между словами в предложении. В отличие от статистических методов анализа, лингвистические методы позволяют осуществить такой анализ. Реализация таких методов является сложным техническим заданием, на основании чего было принято решение использовать сторонний продукт, для осуществления синтаксического и морфологического анализа (например, [5]).

Лингвистический анализ позволяет получить лемму анализируемого слова, а также определить, какой частью речи оно является.

Выделяемое из текста слово имеет следующий вид:

$$W = \{F_{norm,} F_{form}, P\} \quad (2)$$

где F_{norm} – лемма слова, F_{form} – слово в той форме, в которой оно было выделено из текста, P– идентификатор части речи (табл. 1):

Таблица 1

Идентификаторы частей речи

P	Часть речи	P	Часть речи
1	существительное	7	причастие
2	глагол	8	деепричастие
3	прилагательное	9	предлог
4	наречие	10	союз
5	числительное	11	Частица
6	местоимение	12	вводное слово

Выделение леммы F_{norm} является принципиально важным моментом. При статистической обработки множества слов, необходимо определить количество вхождений слова в обрабатываемый текст. Например, в анализируемом тексте можно встретить следующие вариации: «синтаксическим», «синтаксического», «синтаксическом». Все эти слова описывают одну и ту же сущность и имеют одинаковую лемму - «синтаксический».

Вторым важным преимуществом лингвистического разбора является возможность определения части речи. Глаголы описывают действие, а не некоторую сущность предметной области, то есть не могут быть термином. Наречия и прилагательные описывают свойства некоторых объектов предметной области, однако сами по себе не называют их. Наиболее информативной частью предложения являются существительные. Существительные, как правило, являются подлежащими, дополнениями и несут основную смысловую нагрузку. Именно существительные представляют наибольший интерес с точки зрения выделения терминов предметной области.

Как было отмечено ранее, термином может являться не только отдельное слово, но и словосочетание. Во многих случаях именно словосочетание следует считать термином, а не отдельное слово.

Существительное и связанная с ним часть речи называется именной группой. Именной группой также может быть местоимение, заменяющее существительное, и связанная с ним другая часть речи. Однако, именная группа с местоимением не может являться термином, так как местоимение лишь указывает на предмет или лицо, но не называет их. Поэтому, для построения словаря предметной области необходимо выделять именные группы с существительными. Именная группа представляет собой два взаимосвязанных слова:

$$G = \{\{F^1_{norm,} F^1_{form}, P_1\}, \{\{F^2_{norm,} F^2_{form}, P_2\}, L\} \quad (3)$$

где F^1_{norm} и F^2_{norm} - леммы слов, образующих именную группу, F^1_{form} и F^2_{form} - слова именной группы в форме, выделенной из текста, P_1 и P_2 – идентификаторы частей речи слов, входящих в именную группу (см. табл. 1), L – тип связи между словами в именной группе.

Например, для словосочетаний «синтаксический разбор» и «семантический разбор», если опираться лишь на выделение существительных из анализируемого текста, возможна ситуация, когда эти словосочетания будут опознаны как один и тот же термин – будет выделено существительное «разбор», а прилагательное, описывающее тип разбора - «синтаксический» или «семантический» не будет учтено.

Наиболее информативными являются именные группы вида «существительное-прилагательное». Таким образом, именные группы, подлежащие выделению, имеют вид:

$$G = \{\{F^1_{norm,} F^1_{form}, P_1\}, \{\{F^2_{norm,} F^2_{form}, P_2\}, \{1,3\}\} \quad (4)$$

$$G = \{\{F^1_{norm,} F^1_{form}, P_1\}, \{\{F^2_{norm,} F^2_{form}, P_2\}, \{3,1\}\}$$

4. Обработка результатов выделения терминов

В результате синтаксического и морфологического анализа текста, выделаются все слова и именные группы, которые были обнаружены в тексте. Очевидно, что необходимо определить критерий, по которому термины будут заноситься в словарь предметной области.

Количество вхождений получаемых на выходе синтаксического анализатора слов и именных групп в текст не может быть использовано как критерий вхождения в словарь, так как отражает данные только по одному тексту.

Например, в тексте T_1 некоторая именная группа G_1 встретилась 100 раз, а в тексте T_2 именная группа G_2 встретилась 50 раз. Однако величина «количество вхождений» никак не связана с объемом исследуемого текста, т.е. количеством слов и именных групп. Таким образом, текст T_2 может содержать в общей сложности 1000 именных групп, и тогда G_2 составляет 5% от всех встреченных групп, в то время как T_1 может быть намного более объемным и содержать более 10000 именных групп, т.е. G_1 составляет не более 1% от количества ИГ.

Универсальной частотной характеристикой можно принять величину $f_{сл}$ (1), положив n_f = 1000 для слов и 100 – для именных групп (т.к. количество именных групп в текстах невелико относительно количества слов).

Определение вхождения слова или именной группы в словарь предметной области происходит в два этапа:

- предварительное определение границ вхождения
- экспертное определение границ вхождения

Эксперименты показали, что предварительное определение можно выполнить путем сортировки кортежей выделенных слов и именных групп по убыванию их частот и выделению первых 10% элементов данных кортежей.

Предварительно сформированные списки слов и именных групп предоставляются на ревизию эксперту, который определяет, попали ли все важные термины данной предметной области, выделенные из некоторого текста, в словарь предметной области. Эксперт указывает слова, которые, должны или не должны присутствовать в конечном словаре предметной области. Используя частоту указанных экспертом слов, устанавливается новый критерий вхождения слова или именной группы в словарь.

5. Построение словаря предметной области

Процесс формирования конечного словаря предметной области можно рассмотреть с двух позиций:

- формировать словарь предметной области для каждого текста, а общий словарь предметной области будет являться их композицией.

- формировать общий массив слов и именных групп для всех проанализированных текстов и затем формировать словарь предметной области (возможна потеря важных терминов предметной области в связи с неравномерным распределением слов по анализируемым текстам).

Процесс построения словаря предметной области состоит в последовательной обработке некоторого множества текстов $T_{от}$ и выделения из него слов и именных групп, являющихся терминами.

Множество отобранных текстов $T_{от}$ определяется экспертом из некоторого множества исходных текстов $T_{ис}$.

Множество $T_{от}$ может содержать как единичный текст, так и сотни различных документов. Необходимо разработать критерий готовности словаря предметной области.

Мощность множества $|T_{от}| = n$. При последовательной обработке текстов из заданного множества, на каждой итерации обрабатывается очередной текст T_i из $T_{от}$. Для каждого T_i создается пара кортежей $M_{сл}$ i-ое и $M_{иг}$ i-ое, содержащих выделенные из i-го текста слова и именные группы соответственно. Полученные кортежи включаются в $M_{сл.общ}$ и $M_{иг.общ.}$ $M_{сл.общ}$ – кортеж, содержащий выделенные слова всех обработанных текстов, $M_{иг.общ}$, соответственно, содержит все именные группы.

Таким образом, после обработки каждого текста словарь предметной области увеличивается. Процесс его формирования можно остановить после обработки j-го текста, в результате которой прироста количества терминов не произойдет.

6. Построение множества фреймов

При выполнении семантического анализа текстов из предметной области, для которой создан словарь терминов V, задача построения модели знаний значительно упрощается по сравнению с [1]. Более того, появляется возможность провести более качественный анализ, добавляя в модель знаний только факты, специфичные для предметной области. Вместо того, чтобы выделять из текста понятия, в процессе анализа необходимо лишь последовательно выявлять термины W_i, присутствующие в V. В результате заполнения слотов фрейма F_i, соответствующего W_i, модель знаний последовательно пополняется связями между фреймами.

7. Выводы

Разработан метод построения словарей предметных областей для определения множества фреймов, определяющих модель знаний. Проведенные исследования показали, что построение словаря нужно разбить на несколько этапов:

- определение набора текстов для формирования словаря предметной области
- выделение терминов из текста
- статистическая обработка выделенных терминов

Были сформированы рекомендации по отбору текстов экспертом, учитывающие возможную неоднородность предметной области, а также разработана модель весовых коэффициентов для определения степени принадлежности к предметной области.

Проанализировав структуру предложения в русском языке, а также роль частей речи, был сделан вывод, что наиболее информативными словами в предложении являются существительные и именные группы, образованные с их использованием. Именно они составляют основу терминологической базы предметной области и должны быть включены в словарь.

Сформулированы критерии попадания слова и именной группы в словарь предметной области, а также определен признак завершения построения словаря.

Предложенный метод может быть использован в различных приложениях: в информационных системах, для построения справочных и поисковых систем, в системах автоматизированного обучения, при анализе текстов на естественных языках.

Литература

1. О.Б.Кунгурцев, С.М.Бородавкін. Застосування мереж фреймів для побудови моделі вилучення фактів з текстів на природній мові // Искусственный интеллект, 2009. – №4.– С. 202 - 207.
2. А.Б.Кунгурцев, И.В.Барыкина. Формирование словаря предметной области // Искусственный интеллект, 2006. - №1.
3. С.А.Шаров. Частотный словарь русского языка. – РосНИИ. – 2001
4. George K. Zipf. Human Behavior and the Principle of Least-Effort. – Addison-Wesley. – 1949
5. http://cs.isa.ru:10000/dwarf/. Программный пакет синтаксического разбора и машинного перевода.

УДК 004.3:004.492

ПРОГНОЗИРОВАНИЕ ГОДОВОГО РОСТА ТРАФИКА ИНТЕРНЕТ, НА ОСНОВЕ ПОДОБРАННОЙ МОДЕЛИ ВРЕМЕННОГО РЯДА

В.В. Шкарупило*
Контактний тел.: 066-129-73-45
E-mail: vadshkar@yandex.ru

К.Н. Касьян
Кандидат технических наук, доцент*
Кафедра компьютерных систем и сетей
*Запорожский национальный технический университет
ул. Жуковского, 64, г. Запорожье, 69063

Наводиться стислий опис глобальної інфраструктури моніторингу Internet – ATLAS. Проаналізовано характер сучасного Internet-трафіку та здійснене прогнозування його річного зростання

Ключові слова: ATLAS, ISP, AS, Comcast, SPSS, AR1

Приведено краткое описание глобальной инфраструктуры мониторинга Internet – ATLAS. Проанализирован характер современного Internet-трафика и выполнен прогноз его годового роста

Ключевые слова: ATLAS, ISP, AS, Comcast, SPSS, AR1

This article represents the short description of ATLAS global Internet infrastructure. The major findings of modern Internet-traffic are analyzed and the forecast of its annual growth is made

Key words: ATLAS, ISP, AS, Comcast, SPSS, AR1

1. Введение

На сегодняшний день наблюдаются существенные структурные изменения характера Internet-трафика. Хотелось бы особо акцентировать внимание на следующих ключевых аспектах (приведенные данные актуальны на четвертый квартал 2009 года):

- 150 автономных систем (AS) агрегируют в себе более 50% всего трафика глобальной сети Internet;
- около 30 мировых корпораций генерируют 30% всего трафика;
- подавляющий объем трафика циркулирует между клиентом и сервером с данными: за год (с 2008 по 2009 год) относительная доля контентного трафика в общем объеме трафика выросла с 35 до 55%; такой стремительный рост дает все основания полагать, что именно вышеназванный тип трафика следует рассматривать в качестве центрального элемента, с целью анализа и прогнозирования;
- значительно снизилась стоимость услуг поставщиков Internet-сервисов (ISP), что обусловлено, прежде всего, высоким уровнем конкуренции, стре-