ФОРМИРОВАНИЕ ТЕМАТИЧЕСКИХ ЗНАНИЙ НА ОСНОВЕ АНАЛИЗА ЕЯ ТЕКСТОВ СЕТИ ИНТЕРНЕТ

В.П. Гладун
Институт кибернетики им.В.М. Глушкова НАН Украины
glad@aduis.kiev.ua

Н.Д. Ващенко
Институт кибернетики им.В.М. Глушкова НАН Украины
glad@aduis.kiev.ua

В.Ю. Величко
Институт кибернетики им.В.М. Глушкова НАН Украины
glad@aduis.kiev.ua

А.Е. Ткачев
Институт кибернетики им.В.М. Глушкова НАН Украины
glad@aduis.kiev.ua

Ключевые слова: семантический анализ, информационный поиск.

1. Цели и базовые идеи

Среди различных вариантов практического использования хранилищ текстовой информации превалирует потребность выделения информации, обладающей тематическим единством. Это – потребность ученого, журналиста, политика, чиновника, писателя, студента. Обычно тема возникает в виде одного или нескольких понятий, некоторой начальной ситуации, имеющей ряд незаполненных валентностей и ситуационных ролей, которые служат ориентирами для поиска новой релевантной информации. Новая информация порождает “ростки” новых направлений поиска. Этот сложный, иногда психологически мучительный творческий процесс нуждается в автоматизированной поддержке. Тематический поиск требует кропотливой роботы с текстами, хранящимися в библиотеках, архивах, Интернете, текстовых базах данных. Трудность этой работы состоит, в частности, в том, что чаще приходится отбирать не целые тексты, а релевантные теме фрагменты текстов. Содержание многих текстов является переплетением ряда тем. Возникает проблема поиска внутри текстовых документов фрагментов, релевантных заданной теме.

В статье рассматриваются методы, программные средства и результаты отбора тематической текстовой информации. Исследования, представленные в статье, продолжают работы, опубликованные в [1-3].

Решение проблемы отбора текстовых фрагментов по заданной теме связывает в один узел следующие действия:

1) выделение текстов или фрагментов текстов, релевантных исследуемой теме;
2) выделение из релевантной информации наиболее важной, в первую очередь такой, которая определяет и связывает наиболее существенную терминологию темы;
3) представление выбранной информации в удобной для пользователя форме.

В основе реализации указанных действий лежат следующие идеи:

1) ориентировать методику отбора тематической текстовой информации на Интернет, как на наиболее полное хранилище текстовых данных;
2) для отбора релевантной информации сочетать поиск по ключам с семантическим анализом ЕЯ-текстов;
3) использовать семантические критерии отбора наиболее важной тематической информации;
4) организовать автоматический циклический процесс формирования ключевых слов таким образом, чтобы как можно полнее раскрыть исследуемую тему через определения и связи терминов.

2. Методика

Начальный этап отбора тематической информации состоит в поиске в Интернете текстовых документов по заданному ключу. Существующие методы информационного поиска в Интернете выдают много ненужной пользователю "мусорной" информации, фильтрация которой занимает слишком много времени. Выход из положения состоит в использовании в этих целях семантических критериев, обеспечивающих отбор наиболее существенных характеристик понятий, относительно которых собирается информация.

Предлагаемый метод основан на предположении, что наиболее важная пользовательская информация содержится в ядерных конструкциях предложений. Термин "ядерные конструкции" используется в трансформационной грамматике для обозначения простого базового суждения, путем трансформации которого формируется предложение в целом. В данном случае ядерной конструкцией служит предложение, состоящее из подлежащего, сказуемого и соединяющей их связки.

Метод представляет собой циклически повторяющуюся последовательность следующих операций:

1. Отбор заданного количества (параметр) текстов по ключу. Набор используемых поисковых систем неограничен. В настоящее время имеется возможность использования следующих поисковых систем: Yandex, Rambler, Meta - Ukraine, Aport, Google.
2. Выделение в найденных текстах предложений, содержащих заданный ключ.
3. Отбор во множестве предложений, выделенных в п.2, предложений, содержащих ядерные конструкции. Для выполнения п.3 используется естественно-языковый семантический анализатор.
4. Формирование *n*-шаговых расширений ядер выделенных предложений. *n*-шаговым расширением ядра называется часть предложения содержащая его ядро, а также слова, связанные в дереве зависимостей с элементами ядра путями, длина которых не превышает *n*. *n* является параметром, задаваемым пользователем.

   П.4 выполняется на основе семантического анализа предложения.
5. Выделение в множестве предложений, отобранных в п.3, таких предложений, в которых *n*-шаговые расширения ядер содержат заданный ключ.
6. Формирование нового ключа на основе анализа семантических представлений ранее отобранных предложений. Переход к п.1.

Первоначальное ключевое слово задается пользователем. Новые ключи на последующих циклах алгоритма выбираются из числа терминов – знаменательных слов, употребляемых лишь в пределах отдельных предметных областей. Термины отмечаются в словаре.

При выборе нового ключа учитывается степень его релевантности заданной теме, определяемая по результатам семантического анализа предложений. В качестве ключа на

следующем цикле алгоритма выбирается неиспользованный ранее термин с наибольшим коэффициентом релевантности.

После выбора нового ключа действия 1- 6 повторяются.

3. Семантический анализ

Основной операцией семантического анализа естественно-языковых текстов является распознавание синтаксических и семантических отношений, связывающих слова текста. Распознавание отношений осуществляется на основе их описаний (моделей). Такого рода модели обязательно присутствуют во всех методах анализа, хотя не всегда явно. В большинстве методов анализа процессу распознавания отношений предшествует перевод исходного естественно-языкового представления распознаваемых объектов (отношений) в язык категорий традиционной грамматики (число, род, падеж, время и т.д.). Правила распознавания синтаксических и семантических отношений оперируют грамматическими описаниями слов. Привязка к грамматическим описаниям элементов текста влечет следующие недостатки: разнородность процессов обработки отдельных слов и словосочетаний; громоздкость процессов обработки; сложность адаптации к изменениям лексики и предметной области пользователя; трудоемкость разработки. Между тем, переход к грамматическим описаниям не является обязательным условием для выполнения анализа естественно-языковых текстов. Информация, необходимая для распознавания синтаксических и семантических отношений, содержится непосредственно в тексте. Доказательством тому служат “человеческие” процессы анализа естественно-языковых текстов, не связанные с грамматическими категориями и правилами. Поэтому правомочен другой подход, основанный на использовании соответствий между отношениями и средствами их выражения в естественно-языковых текстах. Распознавание синтаксических и семантических связей между знаменательными словами осуществляется путем анализа сочетаний флексий и предлогов, без использования категорий и правил традиционной грамматики. В силу своих принципиальных особенностей такой подход позволяет исключить названные выше недостатки.

Модели отношений, в которых для распознавания синтаксических и семантических отношений используются элементы естественно-языковых текстов, назовем *лексическими моделями отношений*. Алгоритм семантического анализа естественно-языковых предложений на основе лексических моделей отношений описан в [1-3].

4.Реализация и результаты

В составе программного комплекса, реализующего процессы формирования тематических знаний, выделены программы, осуществляющие следующие действия:

1. Отбор в Интернете текстовых фрагментов, содержащих заданный ключ.
2. Формирование семантических представлений предложений (лингвистический процессор).
3. Отбор предложений, релевантных теме, на основе анализа семантических представлений предложений.
4. Выбор нового ключа.

К настоящему времени реализован вариант комплекса для русского языка. В результате работы комплекса формируется текст, состоящий из отдельных предложений, релевантных теме, обозначенной исходным ключом, который задан пользователем. Для каждого предложения указывается адрес документа, из которого оно выбрано. Совокупность предложений, отобранных из одного документа, позволяет сформировать представление о

его тематической релевантности в целом. Высокий уровень релевантности документа может побудить пользователя выбрать этот документ для детального изучения. Совокупность всех отобранных предложений раскрывает исследуемую тему в целом. Степень полноты выделенной информации по теме зависит от эффективности используемой поисковой машины и количества выбранных в Интернете текстов. Множество предложений, отобранных комплексом на основе тематического анализа, хорошо коррелирует с результатом “ручного” отбора “полезных” предложений конечным пользователем. При этом достигается высокая степень отсева информации, ненужной пользователю.

5.Заключение

Описанный метод тематического отбора информации может быть применен для поиска информации не только в Интернете, но и в любых базах текстовых данных. Мы также рассматриваем его, как инструмент создания онтологий. Достоинствами метода является эффективная фильтрация информации по критериям релевантности заданной теме, что достигается за счет применения семантического анализа предложений и циклической организации процесса отбора с автоматическим выбором нового ключа на каждом цикле. Метод допускает сравнительно простую адаптацию к изменениям языка текстов.

Литература

1. Гладун В.П. Процессы формирования новых знаний. – София: СД "Педагогь". 1994г. – 192с.
2. Гладун В.П. Планирование решений. –Киев: Наукова думка, 1987. –168 с.
3. Гладун В.П. Естественный язык в целенаправленных системах. //Диалог’2000. Прикладные проблемы. 2000, с.99-102.

FORMATION OF THEMATIC KNOWLEDGE ON THE BASIS OF ANALYSES OF NL-TEXTS FROM THE INTERNET

V.Gladun, N.Vashchenko, V.Velichko, A.Tkachev

Among various variants of practical use of storehouses for the textual information the necessity to find the information having thematic unity prevails. The paper deals with methods of choice in the INTERNET of natural-language textual fragments that are relevant to a given theme. Existing methods of information search in the INTERNET give out a lot of unnecessary for a user “garbage” information which filtration takes too much time. The way out consists in use of the semantic criteria. Therefore the semantic analysis of sentences is performed. Recognition of syntactic and semantic connections between words of the text is carried out by the analysis of combinations of inflections and prepositions, without use of categories and rules of traditional grammar. Choice in the INTERNET of the thematic information is organized cyclically with automatic forming of a new key at every cycle when addressing to the INTERNET. The created program complex implements described methods for Russian language. The methods allow comparatively simple adaptation to changes of a text language.