УДК 519.7

Н.В. Воронков
СП ЗАО «НаучСофт/IMC», г. Минск, Беларусь, voronkov@scnsoft.com,

Реферирование как задача инженерии знаний

В статье описывается единый процесс автоматического построения реферата текстовых документов применительно к основным элементам базы знаний, а также приводится описание промышленной системы, решающей указанную задачу.

Введение

Понятие «управление знаниями» (УЗ) появилось в середине 90-х годов в крупных корпорациях, для которых проблемы обработки информации приобрели особую остроту и стали критическими. Вскоре оказалось очевидным, что основным «узким местом» здесь является работа (прежде всего сохранение, поиск, тиражирование) со знаниями, накопленными специалистами компании, так как именно этот ресурс обеспечивает ей преимущество перед конкурентами. Зачастую информации в компании накоплено гораздо больше, чем она способна оперативно переработать. При этом часто одна часть предприятия дублирует работу другой только потому, что невозможно найти и использовать знания, накопленные в соседних подразделениях.

Управление знаниями можно рассматривать как направление в менеджменте (стратегия, обеспечивающая интегрированный подход к созданию, организации, использованию и увеличению интеллектуальных и информационных ресурсов предприятия) и как направление в информатике (поддержка процессов создания, распространения, обработки и использования знаний) [1].

Безусловно, разработка таких средств ориентирована на определенный подход и модель (язык) описания данных и знаний. В этом плане все большую популярность в последнее время получают онтологии. Онтология – это точная спецификация некоторой предметной области (ПО), которая включает в себя словарь терминов ПО и множество связей (типа элемент-класс, часть-целое и т.д.), которые описывают, как эти термины соотносятся между собой. Фактически это иерархический понятийный скелет предметной области.

В настоящее время практически все накопленные знания доступны в виде различного рода документов (текстов) в печатной или электронной форме. Поэтому роль автоматической обработки текстов очень высока и продолжает расти. Часто в процессе работы с онтологиями экспертам необходимы ссылки на первоисточники, из которых выделены те или иные знания. Как правило, этими источниками являются тексты в электронном виде. Однако количество этих ссылок может быть очень большим, и время, затрачиваемое на их обработку, может стать серьезным препятствием для работы с онтологическими системами. Возникает необходимость сократить объем информации, который должен быть обработан экспертом. В связи с этим, в частности, появляется задача реферирования текстов.

В общем случае под рефератом понимается связное изложение основных идей текста. Реферирование можно рассматривать и в контексте классических элементов знаний, выделяемых системами их обработки, т.е. как систему, ориентированную на работу с объектами, фактами и закономерностями предметной области. При этом системы реферирования должны иметь механизм выделения указанных элементов, а также уметь упорядочивать их в соответствии с уровнем релевантности к документу и выделять фрагменты текста, описывающие эти элементы. Это, однако, не означает, что речь идет о построении отдельных рефератов, ориентированных на объекты, факты и закономерности ПО. Имеется в виду создание единого процесса реферирования, оперирующего статистическими, лингвистическими, в том числе семантическими, и эвристическими составляющими текстового документа.

Общая схема выделения основных элементов базы знаний

В качестве общей схемы выделения элементов базы знаний из текстового документа можно предложить следующую

1. Преформатирование. Существует много различных форматов документов, и поэтому для упрощения процесса их обработки возникает необходимость конвертации их в некоторый формат, который должен быть удобен для обработки, а также максимально сохранять стилистическую и структурную разметку документов. Кроме того, на данном этапе осуществляется разбиение текста на параграфы, выделение заголовков и подзаголовков, выделение разделов текстов. Также происходит фильтрация вспомогательного текста (текстов кнопок, меню, скриптов и т.д.).

2. Лингвистический (лексический, лексико-грамматический, синтаксический и семантический) анализ текста, в результате которого определяются лексико-грамматические классы (теги) его слов, строится синтаксическое дерево каждой фразы, выделяются синтаксические отношения типа глагольных групп (отношение, связывающее именные группы и их атрибуты посредством глагола, представляющее, по сути, законченную смысловую единицу), распознаются объекты и семантические отношения между ними типа С-А-О (субъект-акция-объект) и отношения типа Причина-Следствие – между самими САО-тройками [2]. Понятно, что САО-тройка соответствует такому классическому элементу знаний, как факт, а отношение Причина-Следствие – закономерности предметной области.

3. Обработка лексико-статистической информации. На этом этапе происходит накопление статистических весов информативных слов текста с целью выделения наиболее значимых слов текста. Причем статистика должна собираться с использованием как отдельных слов, так и объектов, фактов и закономерностей ПО. Для эффективной работы алгоритма выделен ряд тегов так называемых информативных слов. Это существительные, прилагательные, глаголы, наречия, имена собственные и т.д. К неинформативным словам относятся такие части речи, как предлоги, артикли, числительные и некоторые другие. Веса начисляются только значимым словам текста. Для улучшения качества работы статистического алгоритма используется ряд коэффициентов, повышающих веса слов, например, в зависимости от того, является ли слово частью именной группы либо частью семантического отношения САО. Также словам начисляются различные дополнительные веса в соответствии с тем, в каком поле семантического отношения они встречаются, например в

субъекте, акции или объекте, а также являются ли они частью заголовка обрабатываемого документа и т.д.. После обработки всего документа происходит нормализация весов слов таким образом, чтобы веса всех слов были в промежутке от 0 до 1.

Реферирование на уровне объектов ПО

В данном случае речь идет о построении так называемого topic-ориентированного реферата, который позволяет выделить наиболее важные объекты (темы) документов на английском языке, а также построить их иерархическое дерево и получить фрагменты предложений, наиболее релевантные интересующим пользователя темам [3]. В качестве предварительной обработки текста используются шаги 1 – 3, описанные в предыдущем разделе.

Таким образом, на вход модуля собственно реферирования поступает уже лингвистически и статистически обработанный текст. В частности, в каждом из его предложений выделены именные группы, которые являются базовыми элементами процедуры распознавания тем текстового документа. С этой целью каждая именная группа проходит следующие основные преобразования:

- разбиение по союзам «and» и «or» (например, «a device and technique» -> «device», «technique»; «solid and liquid phases» -> «solid phases», «liquid phases»);
- фильтрация неинформативных слов именных групп по лексико-грамматическим тегам (таковыми являются, например, артикли, числительные и т.д.);
- фильтрация неинформативных слов по словарю (неинформативными являются, например, слова «said», «above», «similar», «following» и т.д.);
- так называемая трансформация именной группы по «of» (например, «producers of hazardous materials» -> «hazardous materials producer»; «the end user of the textile» -> «textile end user»).

Затем осуществляется фильтрация неинформативных именных групп в целом. Это производится в 2 этапа на основе заранее заданных шаблонов (паттернов).

На первом этапе фильтруются неинформативные по тегам именные группы, т.е. если именная группа состоит только из слов определенных тегов (стоп-тегов), а также слова содержат определенные символы, то такие именные группы считаются неинформативными, например:

- 6_CD %_NNUS less_JJR
- 185_CD .degree._NNU C._NNU
- FIG._NN
- и т.д.

На втором этапе фильтруются неинформативные по лексическому составу именные группы. Например, в соответствии с паттерном «step + буква» будут отфильтрованы именные группы вида: «step A», «step B» и т.д.

Если именная группа не отфильтрована в результате работы данных фильтров, то она добавляется в список информативных именных групп документа.

После обработки всех предложений текста получается общий список информативных именных групп документа. Далее, каждой именной группе предложения начисляется статистический вес в соответствии с весом слов, полученном в результате работы статистического алгоритма (шаг 3 предыдущего раздела). Вес именной группы берется равным среднему арифметическому весов входящих в нее информа-

тивных слов. Затем производится отсечение некоторого числа именных групп с малым весом, а именно оставляется некоторое число самых «тяжелых» именных групп так, чтобы их общий вес не превосходил вес, равный некоторому проценту от общего веса всех именных групп. В частности, на основании тщательно организованного тестирования в промышленном варианте системы за пороговое значение был взят показатель 80 %.

Все именные группы, оставшиеся после описанных выше преобразований, будем называть темами. Далее, для полученных тем строится иерархическое дерево.

Для этого все темы сортируются по возрастанию количества входящих в них информативных слов, а в пределах одной длины – по возрастанию их веса. Затем, проходя с конца полученного списка к началу, для каждого элемента ищется наиболее подходящий «отец» – он целиком содержится в «сыне», и его длина и вес максимальны. Для проверки включения тем друг в друга требуется O(m1 + m2) операций, где m1 и m2 – длины тем в информативных словах. Для поиска лучших «отцов» требуется O(n * n), где n – количество тем в списке. Однако можно предложить модификацию этого алгоритма, которая в худшем случае имеет аналогичную трудоемкость, а, как показывает практика, в среднем дает ускорение в построении иерархического дерева в 5 – 10 раз. Модификация алгоритма состоит в следующем.

1. Всем словам текста присваиваются идентификаторы от 0 до M – 1, где M – количество уникальных канонических форм слов в тексте.
2. Все темы сортируются по возрастанию количества входящих в них информативных слов, а в пределах одной длины – по возрастанию веса.
3. Далее, учитывая, что все слова в пределах текста имеют уникальные идентификаторы, для каждого слова строится список номеров тем (согласно полученному отсортированному списку), в которых оно встретилось.
4. Затем, начиная с конца списка тем, для каждой темы строится список номеров тем (кандидатов на то, чтобы стать «родителем» рассматриваемой темы), в которых встречаются все ее информативные слова.
5. Полученный список номеров тем сортируется по убыванию.
6. Поиск лучшего «отца» производится в порядке полученного списка возможных кандидатов, номера которых меньше номера текущей темы.

После того как дерево построено, все вершины без «отцов» и все «сыновья» каждой вершины сортируются в порядке убывания весов. Таким образом, получено упорядоченное в порядке убывания релевантности иерархическое дерево тем.

Далее, из каждой ветки дерева выбирается главная тема – тема, наиболее часто встречающаяся в тексте (с точностью до информативных слов и разворотом по *and* и *or*).

На следующем шаге происходит выделение предложений, наиболее релевантных полученным темам, и далее – выделение фрагментов этих предложений с учетом тем (в первом случае применяется метод, используемый при построении так называемого классического реферата [4]). Эта процедура осуществляется следующим образом. После того как все предложения получат оценку их релевантности тексту, для каждой темы выбираются предложения, содержащие ее или ее «детей» в иерархическом дереве. Далее, в конец этого списка добавляются предложения, содержащие «родителя» темы и остальных «детей» «родителя». Предложения сортируются по весу тем, а в пределах равных весов тем – по весу предложений и по положению темы в предложении. Затем производится выделение

фрагментов предложений, содержащих темы, с целью исключения нерелевантной информации. «Усечение» предложений происходит на основе глагольных групп, чтобы по возможности избежать «обрыва» связных фраз, описывающих темы:

1) определяется местоположение темы в предложении;
2) ищется глагольная группа, содержащая данную тему;
3) если глагольная группа найдена, то выбирается самая длинная, но не превышающая некоторого заранее заданного порогового значения глагольная группа, содержащая тему; если глагольная группа не найдена, то именная группа расширяется влево и вправо до заранее определенной минимальной длины усеченного предложения и, в случае, если полученные границы находятся внутри именных групп – далее, до их границ;
4) если полученные «усеченные» предложения достаточно близки к границам исходного предложения, то они расширяются до начала и (или) конца исходного предложения.

Реферирование на уровне фактов и закономерностей ПО

В качестве базы для построения реферата, ориентированного на факты и закономерности предметной области, опять-таки рассматривается процесс построения классического реферата путем выделения наиболее информативных фактов, а также оценки через них релевантности предложений по отношению к содержащему их документу, что в конечном счете позволяет определить наиболее информативные предложения, описывающие рассматриваемые элементы базы знаний.

В основе построения классического лежит комбинация лингво-статистического, позиционного и эвристического алгоритмов [5], в результате работы которых производится комплексная оценка предложений для вычисления их релевантности к тексту.

Лингво-статистический алгоритм представляет собой построение реферата посредством оценки статистической важности фактов и закономерностей ПО, выделенных из текста. Поскольку закономерности ПО, по сути, представляют собой 2 факта – причину и следствие, это дает возможность ограничиться вычислением статистической оценки фактов, входящих в текст, а затем провести оценку закономерностей ПО уже на основе имеющейся оценки фактов. В качестве статистической оценки фактов принимается среднее арифметическое весов входящих в САО информативных слов, с учетом того, в каких полях САО-отношения встретились слова (субъект, или акция, или объект и т.д.), а также количества заполненных полей данного отношения. Далее производится статистическая оценка закономерностей ПО, выделенных из предложений, как среднее арифметическое между весовой оценкой составляющих их причин и следствий с последующим умножением на некоторый «поправочный» коэффициент, учитывающий, что закономерности ПО являются более информативными элементами базы знаний, чем факты. Далее происходит нормализация весовых оценок так, чтобы их значения лежали в интервале от 0 до 1. Весовая оценка предложений в соответствии с данным методом получается как максимум из весов фактов и закономерностей ПО, входящих в предложение.

Позиционный метод позволяет учесть местоположения предложений в тексте, а также учесть расположение фактов и закономерностей ПО внутри предложений. Каждый раздел документа имеет некоторый коэффициент «важности» этого раздела

для документа. Поэтому в процессе обработки документа для каждого слова происходит определение самого важного раздела, в котором это слово встретилось, и после обработки всего документа вес слова умножается на коэффициент значимости этого раздела. Также при получении весовой оценки релевантности предложения тексту учитывается, что наиболее важные предложения находятся ближе к началу/концу текста, к началу/концу параграфа [6].

Из-за субъективизма человеческих оценок разные эксперты часто не могут дать одинаковой оценки релевантности предложений. Также не всегда эксперты могут прийти к согласию, какая из частей предложения является определяющей с точки зрения ее смысла. Например, рассмотрим предложение: «Abbott Laboratories is a global, diversified health care company devoted to the discovery, development, manufacture and marketing of pharmaceutical, diagnostic, nutritional and hospital products». В нем описывается продукция, которая производится компанией. Но с другой стороны, эта же информация указывает на сегмент рынка, на котором представлена фирма, и не указываются конкретные марки продуктов, которые выпускаются фирмой. То есть это предложение может быть отнесено как к общей информации о компании (health care company), так и к информации, продукции и услугах, предоставляемых компанией (pharmaceutical, diagnostic, nutritional and hospital products). Аналогично предложение «ATI is a public company whose shares trade on the Toronto Stock Exchange and NASDAQ.» можно отнести как к общей информации о компании (ATI is a public company), так и к информации о ее финансовой активности (shares trade on the Toronto Stock Exchange and NASDAQ). В связи с этим и возникает необходимость использования некоторых эвристических оценок при выделении информативных предложений. В основе алгоритма, дающего эвристическую оценку, используется метод так называемых «слов-подсказок», имеющий гибкий язык правил, позволяющий оперировать лексическими единицами, синтаксическими и семантическими отношениями, выделенными в тексте, и т.д. При использовании этого метода решение о включении в реферат или исключении из него определённых предложений принимается на основании того, удовлетворяют ли они соответствующим шаблонам (паттернам).

Паттерны могут использоваться с различными целями, например:

- для выделения наиболее информативных предложений текста;
- для структуризации текста в соответствии с некоторым заранее заданным набором полей;
- для удаления вводных частей предложений и т.д.

Каждому правилу часто ставится в соответствие вес, отражающий информативность предложений, выделяемых данным правилом.

В первом случае паттерны используются при построении так называемого классического, однополевого реферата.

Во втором случае паттерны группируются по полям и выполняют две функции: определение принадлежности предложения к одному из полей и определение информативности предложения в пределах одного поля. Предложение относится к тому из полей, для которого «сработал» паттерн с наибольшим весом. В случае если «срабатывает» несколько паттернов из разных полей с одинаковым весом, предложение относится к тому полю, приоритет которого выше.

В итоге все предложения текста получают комплексную оценку, основанную на результатах работы лингво-статистического, позиционного и эвристичес-

кого алгоритмов с учетом некоторых заранее заданных коэффициентов, отражающих «важность» каждого из этих алгоритмов в результирующей оценке. Это дает возможность выбора наиболее релевантных предложений, описывающих как объекты, так и факты и закономерности ПО.

Заключение

Таким образом, задача реферирования как задача инженерии знаний должна иметь механизм выделения основных составляющих базы знаний: объектов, фактов и закономерностей предметной области (внешнего мира) и включать построение реферата, ориентированного на эти 3 основных элемента. Единообразный подход к выделению наиболее релевантных предложений, описывающих различные элементы базы знаний, позволяет создать единый механизм построения этих типов рефератов.

Литература

1. Гаврилова Т.А. Онтологический подход к управлению знаниями при разработке корпоративных информационных систем. Новости искусственного интеллекта. – 2003. – № 2. – С. 24-30.
2. Batchilo L.S., Sovpel I.V., Tsourikov V.M. Document semantic analysis/selection with knowledge creativity capability utilizing subject-action-object (SAO) structures: US Patent 6,167,370, December, 2000.
3. Voronkov N.V., Sovpel I.V. Automatic topic-oriented summarization // Text Processing and Cognitive Technologies. Paper Collection. – 2002. – № 7. – P. 94-102.
4. Воронков Н.В. Использование эвристических оценок в задаче автоматического реферирования текстов // Мат-лы 1-й междунар. конф. «Информационные системы и технологии». – Мн. – 2002.
5. Batchilo L.S., Sovpel I.V., Tsourikov V.M. Computer based summarization of natural language documents: US Patent Appl. № 20030130837.
6. Advances in Automatic Text Summarization. – The MIT Press, 1999.

М.В. Воронков

Реферування як задача інженерії знань

У статті описується єдиний процес автоматичної побудови реферату текстових документів стосовно до основних елементів бази знань, а також проводиться опис промислової системи, яка розв’язує вказану задачу.

Summarization as a Knowledge Engineering Task

The article presents a unified technique of summarization of text documents as applied to the main elements of knowledge base. An industrial system which solves the above stated task is described as well.

Статья поступила в редакцию 12.07.2004.