УДК 681.82:519.81

З.А. Алисейко, О.В. Канищева

АВТОМАТИЗИРОВАННОЕ ИНДЕКСИРОВАНИЕ ПОЛНОТЕКСТОВЫХ ДОКУМЕНТОВ КЛЮЧЕВЫМИ СЛОВАМИ

Актуальность поставленной задачи. В последнее время в технологии поиска все чаще стали применяться элементы контент-анализа, методологии возникших в конце XIX - начале XX вв. Эта методология, изначально ориентированная на применение в психологии и социологии, сегодня все чаще используется в разного типа автоматизированных системах. Различают количественный и качественный контент-анализ. Если качественный контент-анализ базируется на глубоком лингвистическом и семантическом анализе отдельных предложений и всего текста, то основой количественного контент-анализа являются статистические подходы [1].

В последнее время получили развитие такие направления контент-анализа, как "Data Mining" и "Text Mining", которые предполагают автоматическое выявление нового смысла из текстовых массивов, новых данных, феноменов, фактов - знаний. Все чаще возникают попытки привлечения методов контент-анализа, а точнее Text Mining в реальные поисковые системы. Во многие современные поисковые системы внедрены такие компоненты, как: автоматическая группировка документов, по определенному пользователем классификатору; автоматическое определение новых, не заданных заранее классов, на основе неструктурированных или слабо структурированных документов; ранжирование документов по поисковой релевантности; выявление семантически подобных документов - поиск подобных документов относительно эталона; автоматический анализ и смысловое преобразование запросов пользователей. Однако поисковые системы далеки от совершенства. Многие из них используют лишь количественный контент-анализ, без учета возможных словоформ, также не учитывается тематика текста, по которой можно было бы давать преимущества тем или иным словам, связанным с обсуждаемой в документе темой.

В данной статье будет рассмотрен вопрос об индексирования полнотекстовых документов ключевыми словами с учетом семантического анализа. Он непосредственно связан с вопросом поиска ключевых слов в тексте.

В начале определимся с такими понятиями как индексирование, координатное индексирование, дескриптор и ключевое слово. *Индексирование* – выражение содержания документа и/или смысла информационного запроса на информационно-поисковом языке. *Координатное индексирование* – индексирование, предусматривающее многоаспектное выражение смыслового содержания документа или смыслового содержания информационного запроса множеством ключевых слов или дескрипторов. *Дескриптор* – лексическая единица, выраженная информативным словом (вербально) или кодом и называющаяся именем класса синонимичных или близких по смыслу ключевых слов. *Ключевое слово* – информативное слово, приведенное в стандартной лексикографической форме и используемое для координатного индексирования. Очень часто под понятием дескриптора понимают ключевые слова и словосочетания [1]. В традиционном понимании ключевыми словами называются полнозначные слова, устойчивые сложные и сложносокращенные слова и терминологические словосочетания, несущие смысловую нагрузку в текстах документах. Устойчивое словосочетание – это словосочетание, при изменении или изъятии слов из которого, меняется его смысл.

Индексирование текстов является одной из проблем при создании информационно-поисковых систем, использующих в качестве критериев поиска набор ключевых слов. Вся задача состоит в том, как "правильно" определить этот набор. К сожалению, однозначного ответа на этот вопрос дать нельзя. На сегодняшний день *существует довольно много различных вариантов поиска текстов (или их фрагментов)* по ключевым словам. Конечно, каждый из них имеет свои достоинства и недостатки.

Но сначала выясним проблему, связанную с индексированием текстов. Она состоит в том, что от ключевых слов (индексов) требуется соблюдение двух взаимоисключающих принципов:

- Ключевые слова должны как можно точнее идентифицировать текст.
- Ключевые слова должны как можно более точно отражать содержание (смысл) текста.

Рассмотрим каждый из этих принципов. Предположим, что определенные ключевые слова полностью идентифицируют текст в заданном подмножестве текстов. Из этого автоматически вытекает, что выбраны такие ключевые слова, которые не встречаются ни в каком другом тексте, кроме того, который они определяют. Понятно, что такими словами могут быть только специфические термины, фамилии авторов, названия каких-то малоизвестных фирм и т.п. Определив, таким образом, ключевые слова, пользователь информационно-поисковой системы должен обязательно их помнить. Но, как правило, пользователь не может запомнить какие-то крайне редкие термины и хочет видеть в списке ключевых слов те, которые, по его мнению, отражают смысл текста. Очевидно, что редкие термины далеко не всегда являются центральными в тексте, хотя и полностью его идентифицируют. Отсюда возникает противоречие со вторым принципом.

Рассмотрим теперь другой крайний случай. Пусть ключевые слова полностью отражают смысл текста. Но тогда вероятность получения только какого-то одного требуемого текста сильно снижается, поскольку текстов, сходных по смыслу в заданном подмножестве текстов, может быть несколько. Противоречие с первым принципом. Кроме того, остается неясным, как отобрать те ключевые слова, которые полностью отражают смысл текста.

В общем случае эта проблема однозначно не разрешима, хотя и существуют достаточно эффективные системы поиска (например, поисковые системы в Internet). Однако автоматическое индексирование и поиск ключевых слов в полнотекстовых документах необходимо проводить не только в Internet, но, и в современных библиотеках, которые нарастающими темпами накапливают неструктурированные текстовые ресурсы. Причем объем накопленной текстовой информации может быть таким затруднительным, что задача подготовки их полного библиографического описания становится крайне затруднительной. Очевидна необходимость применения специальных решений, которые позволят работнику библиотеки автоматизировать процесс обработки полнотекстовых документов.

Поиск ключевых слов в полнотекстовых документах. Уже сейчас в Интернете можно найти множество анализаторов текста, которые, обрабатывая текст, находят его ключевые слова. Вот некоторые из них:

– Анализатор текстов – Интернет-приложение, которое занимается поиском ключевых слов (http://rmckttt.pstu.ac.ru/Articles/wm/wm17.html).

– META Tuner -свободно-распространяемый анализатор текста на повторяющиеся слова (http://www.metasset.com/tools/mtuner/download.html).

Но все эти программы имеют ряд недостатков. Большинство из них работает лишь с подсчётом повторений слов, без учета возможных словоформ, т.е. с методами количественного контент-анализа. Также не учитывается тематика текста, по которой можно было бы давать преимущества тем или иным словам, связанным с обсуждаемой в документе темой.

В данной статье авторами предлагается использовать следующий подход к нахождению ключевых слов в русскоязычных документах (см. рисунок).

Необходимо отметить, что роль существительных в данной предметной области гораздо существенней, чем глаголов или прилагательных (логично относить иноязычные слова к существительным, поскольку они определяют объекты реального мира). Число существительных значительно превосходит число глаголов и прилагательных. Из этого следует, что рассматриваемая предметная область представлена в основном понятиями и терминами. Поэтому сначала ищутся имена существительные с наибольшим количеством повторений в тексте. Найденные имена существительные обрабатываются с помощью морфологического анализа (МА). Для анализа русских текстов, как правило, используют алгоритмический МА [2,3]. При этом в словарях хранятся как основы, так и окончания словоформ, которым приписаны необходимая информация. МА осуществляется путем выделения в составе анализируемой словоформы некоторой словарной основы и некоторого словарного окончания. Затем производится сопоставление информации об основе и окончании и получается комплекс морфологической информации ко всей словоформе. Для удобства всех анализа всех окончаний каждой части речи группируются по словоизменительным типам. Например, для морфологического анализа научно-технических текстов можно обойтись шестнадцатью словоизменительными типами имен существительных.

После морфологического анализа, найденные слова приводятся в каноническую форму. Однако не всегда ключевые слова представлены, одним словом. Для поиска ключевых слов в виде словосочетаний необходимо решить ряд следующих вопросов.

Недостатком многих программ является то, что нет поиска ключевых слов следующей конструкции: “прилагательное + существительное”, а также не учитываются случаи замены существительных на местоимения при подсчёте количества повторений слова в тексте. Для решения первой проблемы предлагается выделить грамматические признаки по характерным окончаниям прилагательных и присвоить эти признаки существительным. А для решения второй – хранить все словоформы местоимений в БД словаря, в связи с тем, что в русском языке сравнительно мало местоимений, а их вес в частотной таблице велик. Этим, кроме того, обеспечивается возможность поиска связок местоимений и имен.

Также необходимо использовать небольшую базу со словами, не являющимися ключевыми по определению (союзы, предлоги, частицы…). Она позволит не учитывать при подсчете весов эти слова.

При программной реализации предлагается использовать словарь «Раздел знаний». Этот словарь облегчит поиск ключевых слов в документе, и будет содержать синонимы часто употребляемых терминов в какой-либо предметной области. Перед обработкой программы какого-либо документа библиотекарь сможет выбрать необходимый словарь по нужной ему предметной области.

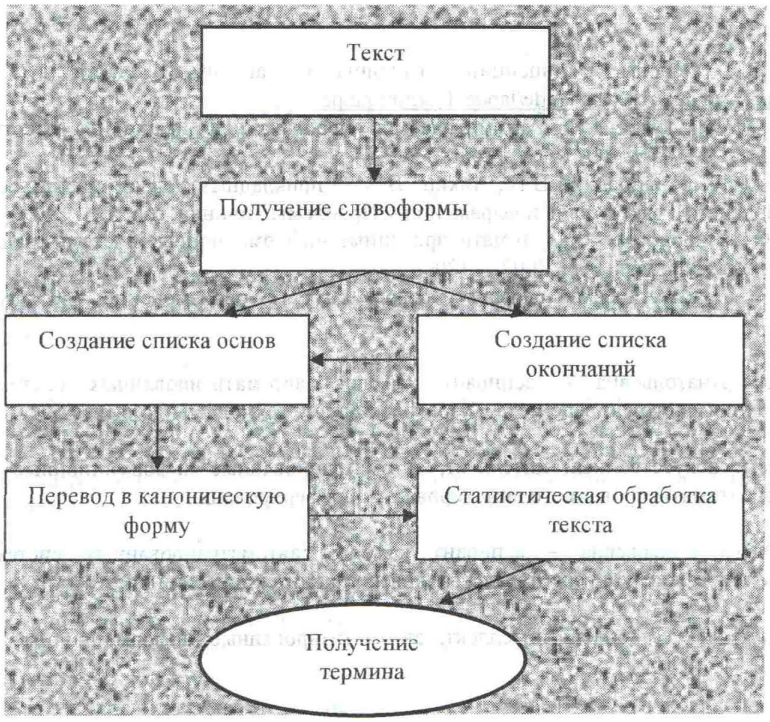

Рис. Этапы получения ключевого слова в полнотекстовом документе

Основные результаты и выводы. Современная лингвистика все шире опирается на данные, получаемые в результате разнообразного применения компьютерного анализа текстов. Несмотря на то, что достаточно много уже достигнуто в области создания моделей искусственного интеллекта, имитирующих восприятие текста человеком, на сегодняшний день более надежными и продуктивными являются методы морфологического, синтаксического и статистического анализа текстов [4]. Эти классические методы обработки текстов, несмотря на то, что применяются в течении нескольких десятилетий нисколько не утратили ценности как с точки зрения прикладных задач, так и научных разработок.

Анализ современных информационных библиотечных систем показал, что значительную пользу для качественного и быстрого накопления библиографических записей полнотекстовых документов может принести программы автоматизированного поиска ключевых слов и словосочетаний. Таким образом, библиотекарь не только будет экономить полезное рабочее время и эффективно пополнять полнотекстовую базу данных, но и качественно, быстро и надежно определять ключевые слова для обрабатываемого документа.

Поскольку Text Mining – это процесс обработки ранее неизвестных нетривиальных практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности, то предложенный в данной работе метод по отношению к библиотечным данным можно квалифицировать как синтез количественного контент-анализа и качественного контент-анализа. Таким образом, представленный метод обработки текстовой информации в АИБС, использующие основные подходы Text Mining, представляют интерес для разработчиков информационно – поисковых, библиотечных и информационных систем широкого назначения.

Достоинствами предложенного метода поиска ключевых слов является

– правильное нахождение необходимой словоформы;

– учитывает случаи замены существительных на местоимения при подсчёте количества повторений слова в тексте (существует некоторая вероятность ошибки, в связи с многообразием и богатством русского языка);

– поиск ключевых слов следующей конструкции: “прилагательное + существительное”.

Особенно удобно использовать данную программу на текстах с большим объёмом, т.к. при таком количестве информации трудно вручную выбрать наиболее важные слова, а, воспользовавшись автоматизированным поиском достаточно всего лишь проверить полученный список ключевых слов и взять самое необходимое на взгляд библиотечного работника.

ЛИТЕРАТУРА:

1. Ландэ Д.В. Основы концепции глубинного анализа текстов (Text Mining) http://download.yandex.ru/class/lande/lande-11-tmining.ppt
2. Бондаренко М.Ф., Осыка А.Ф. Автоматическая обработка информации на естественном языке. Учеб. Пособие. – К. УМК ВО, 1991. – 144с.
3. Бондаренко М.Ф., Рублинецкий В.И., Чикина В.А. О прикладных задачах машинной лингвистики, решаемых подсчетом частот слов и выражений // Проблемы бионики. вып 50 – 1999.
4. Хайрова Н.Ф., Шаронова Н.В. Автоматизированные информационные системы: задачи обработки информации – Х.:Нар. укр. акад.,2002. – 120с.
5. Зализняк А.А. Грамматический словарь русского языка: Словоизменение. - М.: Рус. яз., 1980. – 880 с.

АЛИСЕЙКО Зоя Анатольевна – аспирант кафедры автоматизированных систем управления Национального технического университета «Харьковский политехнический институт».

Научные интересы: искусственный интеллект, автоматизированные информационные библиотечные системы, системы автоматизированного аннотирования и реферирования

КАНИЩЕВА Ольга Валерьевна – аспирант кафедры автоматизированных систем управления Национального технического университета «Харьковский политехнический институт».

Научные интересы: искусственный интеллект, автоматизированные информационные библиотечные системы