УДК 004.912

МЕТОД НАВИГАЦИИ ПО ТЕКСТУ ДОКУМЕНТА С ПОМОЩЬЮ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ЕГО СОДЕРЖИМОГО

А.И. Табарча

Описывается подход, который может быть использован в качестве альтернативы автоматическому реферированию текста. Суть подхода заключается в формировании представлений исходного текста и возможности перемещаться по его содержанию с помощью этих представлений – от общего представления к более конкретному представлению и обратно. Представления формируются на основании методов автоматической обработки текста – статистических методов и поверхностного лингвистического анализа. В работе дано формализованное описание подхода, а также рассмотрена реализация на основе реляционной базы данных.

Ключевые слова: автоматическое реферирование, автоматический анализ текста, автоматическое извлечение терминов, статистические методы.

Введение

Постоянный рост объемов информации снижает эффективность ее обработки традиционными методами. Единственным инструментом, который потенциально может обеспечить охват информационных ресурсов, являются различные программы автоматической обработки текста, реализующие индексирование, аннотирование, реферирование, фрагментирование и другие формы информационного анализа и синтеза [1].

Среди задач автоматической обработки текста можно выделить задачу автоматического реферирования текстов, потребность в решении которой стабильно возрастает. Предполагается, что на основании реферата, составляющего, как правило, 5–30% исходного текста, можно составить обоснованное заключение о первичном документе, затратив значительно меньше усилий на ознакомление с ним [1].

Решения задачи автоматического реферирования четко разбиваются на два направления – квазиреферирование и краткое изложение содержания первичных документов. Квазиреферирование основано на выделении из текста наиболее информативных предложений и формировании из них квазирефератов. Краткое изложение исходного материала основывается на выделении из текста, на основании искусственного интеллекта и с помощью специальных искусственных языков, наиболее существенной информации и порождении нового текста, содержательно обобщающего первичный текст [2].

Проанализировав вышеописанные подходы к решению проблемы стремительного роста производства информации, можно отметить, что более трудоемкий в реализации и поддержке метод краткого изложения содержания дает лучшие результаты, чем более простой в реализации и более универсальный метод квазиреферирования [3].

На практике люди редко приходят к единому мнению относительно наличия необходимой информации в автоматически созданном реферате, вне зависимости от подхода, который был применен. Отсюда можно заключить, что разным людям для формирования представления об исходном документе требуется различная информация из исходного текста. В связи с этим предлагается подход к формированию представлений документа в виде списков сгруппированных терминов, фрагментов предложений, целых предложений, а также фрагментов исходного текста. Подход основывается на методах квазиреферирования, но отличается от обоих вышеописанных методов. Отличительной особенностью метода является то, что пользователю вместо готового реферата предоставляется возможность на основании компактных представлений документа перемещаться по исходному тексту документа от общего представления, отражающего большую часть исходного текста, к подробному представлению меньшей части исходного текста. Таким образом, предполагается обеспечить необходимую гибкость в предоставлении информации, наиболее соответствующей интересам пользователя.

Целью данной работы является разработка системы навигации по тексту документа, которую можно было бы использовать наряду с системами автоматического реферирования. Для достижения цели работы должны быть выполнены следующие задачи:

- детальное описание метода, лежащего в основе работы системы;
- рассмотрение основных вопросов реализации системы.

Описание предлагаемого подхода

Предлагается построение иерархической структуры представлений исходного текста. Каждое представление отражает минимально необходимую информацию для принятия решения об углублении, либо возврате на верхний уровень. Процесс начинается с построения представления верхнего уровня. Далее, в зависимости от выбранного элемента, строятся представления более низких уровней.

Процесс чтения текста можно представить как последовательное распознавание слов, словосочетаний и взаимосвязей между ними. Среди слов текста можно выделить общеупотребительные слова и слова, относящиеся к определенной области знания (термины) [4]. Термином может быть как слово, так и словосочетание. Общеупотребительные слова играют связующую роль. Слова предметной области обозначают объекты. Под объектом понимается любая сущность, понятие или явление. Таким образом, текст можно рассматривать как описание свойств объектов и взаимосвязей между ними. С точки зрения психологии понимание чего-либо основывается на выстраивании связей между объектами. Как следствие, чтобы понять текст, из него нужно выделить описываемые объекты и взаимосвязи между ними.

Если рассматривать текст как описание объектов и взаимосвязей между ними, то для формирования представления верхнего уровня достаточно выделения списка объектов (терминов) из текста. Так как в тексте содержится множество терминов и не все они одинаково явно выражают тематику текста, следует выбрать наиболее значимые термины.

В длинных текстах даже значимых терминов достаточно много, что неудобно для восприятия, поэтому предлагается ввести в описываемую иерархию более высокий уровень представления – совокупность групп терминов. Термины могут быть сгруппированы в соответствии с локальностью их употребления в тексте. Для этого текст может быть разбит на пять-семь частей, в каждой из которых могут быть выделены три-четыре наиболее значимых термина для обозначения самой группы.

Чтобы учесть различную длину исходного текста, процедуру выбора группы можно повторить в зависимости от минимально выбранной значимости терминов, используемых для обозначения групп. При формировании подгруппы для выбранной группы термины, обозначающие группу, исключаются из возможных кандидатов в названия подгрупп, что позволяет получить представление о контексте, в котором используются основные термины. Назначение *представления групп терминов* заключается в отражении локальности использования определенных групп терминов в соответствующих частях текста. Здесь и далее по тексту курсивом выделены названия представлений, предложенные автором.

При выборе группы, не имеющей подгрупп, отображается *представление терминов. Представление терминов* представляет совокупность терминов либо отсортированных, либо выделенных изменением размера в зависимости от значения. Назначение представления терминов заключается в отражении наиболее значимых терминов из выбранной части текста. Для дополнительной выразительности наиболее значимые термины выделены по сравнению с остальными.

При выборе элемента *представления терминов* должно отображаться *представление ассоциированных терминов и фрагментов предложений. Представление ассоциированных терминов и фрагментов предложений* состоит из двух частей. Первая часть – это *представление терминов ассоциированных объекто*в, а вторая список фрагментов ассоциированных предложений.

Ассоциированным термином считается термин, находящийся в одном и том же предложении с термином *представления терминов* и не отделенный каким-либо знаком препинания, обозначающим конец предложения. Под ассоциированным предложением рассматривается предложение, содержащее

термин. Фрагмент ассоциированного предложения – это часть предложения, в которую входит термин, вместе с несколькими словами, находящимися с обеих сторон от него.

Назначение *представления ассоциированных терминов и фрагментов предложений* заключается в отражении наиболее значимых терминов из ассоциированных выбранному. Это показывает наличие некой взаимосвязи между объектами, которую можно уточнить, обратившись к представлению более низкого уровня. Предложение не приводится сразу полностью по нескольким причинам. Во-первых, нужно сконцентрировать внимание на наиболее важной части, а таковой считается часть вокруг выбранного объекта. Во-вторых, этой части может быть достаточно для того, чтобы не читать все предложение.

Выбор ассоциированного термина вызывает *представление фрагментов ассоциированных предложений*. В данном случае под ассоциированными предложениями понимаются предложения, содержащие оба термина, как термин верхнего уровня, так и ассоциированный термин. Отображаемый фрагмент предложения включает в себя оба термина вместе со словами, находящимися между ними.

Выбор предложения вызывает *представление контекста*. *Представление контекста* – это отображение предложений, окружающих выбранное предложение. *Представление контекста* по аналогии с *представлением групп терминов* может быть многоуровневым, т.е. при выборе контекста он расширяется вверх, вниз или и вверх, и вниз в зависимости от выбора пользователя и в соответствии с границами исходного текста. Так как представления последующих уровней контекста включают в себя представления предыдущих уровней контекста, то ранее прочитанные части текста выделяются цветом. Объем каждого последующего *представления контекста* равен удвоенному значению объема предыдущего представления с поправкой на симметричность при расширении контекста и вверх и вниз, и т.д. вплоть до достижения границ исходного текста. Общая схема представлений изображена на рисунке.

Представление групп терминов
– **T1.1,T1.2,T1.3** – обозначение первой группы терминов

Представление терминов
– **T1.1** – первый термин первой группы

Представление ассоциированных терминов и фрагментов предложений
– **AT1.1(T1.1)** – первый термин ассоциированный с термином T1.1

Представление фрагментов ассоциированных предложений
– **ФАП1.1(T1.1, AT1.1)** – первый фрагмент предложения, содержащий T1.1 и AT1.1

Представление ассоциированных предложений
– **П1.1(ФАП1.1)** – полное предложение, включающее в себя ФАП1.1

Представление контекста
– **К1.1(П1.1)** – контекст предложения П1.1

+ **ФАП1.2(T1.1, AT1.1)** – второй фрагмент предложения, содержащий T1.1 и AT1.1
…
+ **ФАП1.N(T1.1, AT1.1)** – последний фрагмент предложения, содержащий T1.1 и AT1.1

+ **AT1.2(T1.1)** – второй термин, ассоциированный с термином T1.1
…
+ **AT1.N(T1.1)** – последний термин, ассоциированный с термином T1.1
……
+ **ФАП1.1(T1.1)** – первый фрагмент предложения, содержащий T1.1
+ **ФАП1.2(T1.1)** – второй фрагмент предложения, содержащий T1.1
…
+ **ФАП1.N(T1.1)** – последний фрагмент предложения, содержащий T1.1

+ **T1.2** – второй термин первой группы
…
+ **T1.N** – последний термин первой группы

+ **T2.1,T2.2,T2.3** – обозначение второй группы терминов
...
+ **TN.1,TN.2,TN.3** – обозначение последней группы терминов

Рисунок. Общая схема представлений

В дополнение следует добавить, что с любого представления есть возможность вернуться обратно, на представление более высокого уровня или же на начальное представление.

Рассмотрение некоторых вопросов реализации на основе реляционной базы данных

На подготовительном этапе происходит определение основных элементов текста, выделение их из текста и сохранение в удобном для произвольного доступа виде, чтобы обеспечить формирование различных выборок.

В данной системе выделяются следующие элементы текста:

- предложения;
- разделители предложений;
- разделители слов;
- слова – любая цифробуквенная последовательность и знак дефиса без разделителей;
- именованные последовательности – инициалы и аббревиатуры.

Морфологический анализ – важный этап предварительной обработки текста, так как для последующего анализа необходима морфологическая информация слов, выделенных из текста.

В качестве хранилища для сохранения результатов анализа выбрана реляционная база данных, так как впоследствии она сможет обеспечить простой и удобный доступ к элементам текста в произвольном виде. База данных состоит из 4 таблиц:

1. для хранения информации о тексте;
2. для хранения предложений текста и дополнительной информации о них;
3. для хранения элементов предложений;
4. для хранения лексем и их морфологических признаков.

Когда исходный текст разобран и сохранен в соответствующих таблицах реляционной базы данных, все необходимые представления могут быть сформированы выполнением запросов к базе данных.

Важным вопросом для первых трех уровней представлений является выделение терминов. На этапе анализа использовались возможности автоматического морфологического анализа, и каждой лексеме были сопоставлены ее грамматические характеристики. На основании этих данных можно выделить все лексемы, которые являются существительными, как термины. В процессе анализа исходного текста, кроме словоформ, выделялись также другие именованные последовательности символов, как, например, инициалы и аббревиатуры. Последние также могут быть использованы как термины. Таким образом, чтобы выбрать из текста объекты, нужно выбрать все существительные, а также инициалы и аббревиатуры.

Также следует заметить, что некоторые термины выражаются в предложении более чем одним словом. Так как большинство словосочетаний ограничивается двумя словами [5], то в данной работе производиться поиск лишь двухсловных словосочетаний. Для поиска словосочетаний используются следующие шаблоны:

- прилагательное + существительное – для описания устойчиво сочетающихся с существительным прилагательных и причастий;
- существительное + существительное – для сочетаний двух существительных;
- глагол + существительное – для устойчивых сочетаний глагола и управляемого им существительного.

Для того чтобы выделить именно устоявшиеся словосочетания, а не случайное совместное употребление слов, при выборе учитывается частота встречаемости словосочетания в тексте.

Следующий немаловажный вопрос в выделении терминов из текста – это определение их значимости. В данной работе под значимостью термина понимается совокупность частоты употребления и длины слова/словосочетания, отображающего объект в тексте.

Рассмотрим этапы формирования одного из запросов (выборка списка наиболее значимых терминов из заданного участка текста).

1. Выборка предложений для заданного текста.
2. Ограничение выборки определенной частью последовательно располагающихся предложений (например, с 40-го по 140-е предложение).
3. Выборка элементов предложений только для заданного набора предложений.
4. Объединение выборки элементов предложений с данными таблицы, содержащей необходимую грамматическую информацию.
5. Применение фильтров различных видов словосочетаний для полученной выборки.
6. Выборка словосочетаний с соответствующей им значимостью; значимость определяется как частота встречаемости, умноженная на сумму средних длин словоформ для лексем выбранного словосочетания.

В тексте, кроме словосочетаний, могут встречаться и однословные термины. Для этого выполняется отдельный запрос, после чего данные обоих запросов объединяются и сортируются в соответствии со значимостью. Для выборки однословных терминов достаточно выполнить этапы 1–4 формирования запроса выборки списка наиболее значимых терминов. После этого следует выбрать все существительные,

инициалы и аббревиатуры с соответствующей им значимостью. Значимость инициалов и аббревиатур, ввиду их менее частого употребления, но большей важности по сравнению с обычными словами, должна считаться особым образом.

Сравнительная оценка с уже существующими подходами

Описанный подход призван решать задачу автоматического реферирования текстов. Решение задачи автоматического реферирования текстов разбивается на два направления – квазиреферирование и краткое изложение содержания первичных документов. Краткое изложение содержания первичных документов, ввиду ресурсоемкости и совершенного другого уровня анализа, некорректно сравнивать с описываемым подходом, хотя результаты предлагаемого подхода могут оказаться интересными, ввиду интерактивности подхода.

Сравним предлагаемый подход с методом квазиреферирования. Квазиреферирование основано на выделении наиболее информативных предложений и формировании из них квазирефератов. На практике люди редко приходят к единому мнению относительно наличия необходимой информации в автоматически созданном реферате. Увеличение размера квазиреферата не всегда решает эту задачу, к тому же это увеличивает время на ознакомление с материалом. Преимущество описываемого метода – в том, что пользователю вместо готового реферата предоставлена возможность на основании компактных представлений перемещаться по исходному тексту документа. Таким образом, обеспечивается необходимая гибкость в предоставлении информации наиболее соответствующей интересам пользователя. Реализация системы на практике подтвердила, что преимущество предлагаемого метода по сравнению с квазиреферированием – это гибкость и интерактивность. В качестве недостатка можно рассмотреть случай, когда квазиреферат сразу содержит информацию, интересующую пользователя, а в предлагаемой системе ее нужно найти, переходя между представлениями текста.

Заключение

Описанный подход основывается на методах квазиреферирования – статистические методы и поверхностный лингвистический анализ. Как следствие, подход обладает все плюсами и минусами методов. Но за счет возможности интерактивного взаимодействия с системой неточности автоматического анализа текста, присущие универсальным методам обработки без глубокого лингвистического анализа, сглаживаются.

Отличительная особенность предлагаемой системы – это возможность формировать из исходного текста представления наиболее соответствующие информационным потребностям пользователя. Система позволяет сформировать общее представление об исходном тексте, но обладает также возможностью формирования представлений, позволяющих перейти от общего представления непосредственно к тексту и обратно, поэтому систему можно также рассматривать как средство для поиска необходимой информации в тексте.

Система особенно полезна для сложных текстов, так как способствует выделению значимых объектов и взаимосвязей между ними, фокусировке внимания именно на важных частях текста.

Систему можно улучшить, разрабатывая новые алгоритмы для выделения терминов и ассоциированных терминов, ее несложно модифицировать для формирования представлений коллекции текстов.

Литература

1. Ландэ Д.В. Поиск знаний в Internet. Профессиональная работа. – М.: Вильямс, 2005. – 272 с.
2. Соловьев В.Д., Добров Б.В., Иванов В.В., Лукашевич Н.В. Онтологии и тезаурусы. Модели, инструменты, приложения: Учебное пособие. – М.: Бином.ЛЗ, 2009. – 173 с.
3. Стернин И.А. Методологические проблемы когнитивной лингвистики: Научное издание. – Воронеж: ВорГУ, 2001. – С. 36–46.
4. Кузнецов И.П., Мацкевич А.Г. Лингвистические и алгоритмические аспекты выделения объектов и связей из предметно-ориентированных текстов // Компьютерная лингвистика и интеллектуальные технологии: Труды Междунар. конф. Диалог'2007. – Бекасово, 2007. – С. 333–342.
5. Браславский П., Соколов Е. Сравнение четырех методов автоматического извлечения двухсловных терминов из текста // Компьютерная лингвистика и интеллектуальные технологии: Труды Междунар. конф. Диалог'2006. – М.: Изд-во РГГУ, 2006. – С. 88–94.

Табарча Александр Иванович – Санкт-Петербургский государственный университет информационных технологий, механики и оптики, аспирант, a.tabarcha@gmail.com