УДК 519.7

Н. В. БОРИСОВА, аспирант НТУ «ХПИ»,
О. В. КАНИЩЕВА, аспирант НТУ «ХПИ»

МОДЕЛИРОВАНИЕ СИНТАКСИЧЕСКОГО АНАЛИЗА В ЗАДАЧАХ АННОТИРОВАНИЯ И РЕФЕРИРОВАНИЯ ПОЛНОТЕКСТОВЫХ ДОКУМЕНТОВ

В статті пропонується модель синтаксичного аналізу у задачах анотування та реферування повнотекстової інформації, з використанням алгебри скінчених предикатів та предикатних операцій. Проведено аналіз існуючих методів анотування та реферування повнотекстової інформації. При анотуванні та реферуванні запропоновано використовувати статистичні методи та розроблену модель разом.

В статье предлагается модель синтаксического анализа в задачах аннотирования и реферирования полнотекстовой информации, основанная на использовании конечных предикатов и предикатных операций. Проведено анализ существующих методов аннотирования и реферирования полнотекстовой информации. При аннотировании и реферировании предложено использовать статистические методы и разработанную модель вместе.

In article the model of the syntactic analysis in problems of annotation and abstracting is offered to the text-through information, based on use of final predicates and predicate operations. It is lead the analysis of existing methods of annotation and abstracting the text-through information. It is offered to use at annotation and abstraction statistical methods and the offered model together.

Введение. Актуальность проблемы аннотирования и реферирования. Искусство реферирования, или составления аннотаций, или кратких

изложений материала, иными словами, извлечения наиболее важных или характерных фрагментов из одного или многих источников информации, стало неотъемлемой частью повседневной жизни.

Задача аннотирования документов является актуальной для любых хранилищ информации: от библиотек до интернет-порталов [1, 2]. Потребности в средствах автоматического реферирования и аннотирования испытывают: корпоративные системы документооборота; поисковые машины и каталоги ресурсов Интернет; автоматизированные информационно-библиотечные системы; каналы вещания; службы рассылки новостей и другие [3]. Аннотирование требуется также и конкретному человеку, например, для быстрого ознакомления с интересующей его публикацией или с подборкой статей по определенной тематике.

Хотя некоторые производители уже сейчас предлагают инструменты для реферирования, объем информации в Сети растет и оперативно получать ее корректные сводки становится все сложнее. Такие инструменты, как функция AutoSummarize в Microsoft Office 97, системы IBM Intelligent Text Miner, Oracle Context и Inxight Summarizer (компонент поискового механизма AltaVista), безусловно, полезны, но их возможности ограничены выделением и выбором оригинальных фрагментов из исходного документа и соединением их в короткий текст [1, 2]. Подготовка же краткого изложения предполагает передачу основной мысли текста, и не обязательно теми же словами.

Текст, полученный путем соединения отрывочных фрагментов, лишен гладкости, его трудно читать. Кроме того, источники информации вовсе не всегда являются текстами, ведь необходимо подготавливать аннотации и на видеозаписи, к примеру, спортивных соревнований, или формировать сводные данные по биржевым таблицам. Перечисленные инструменты реферирования рассчитаны на обработку только текстовой информации. И, наконец, они не могут работать сразу с несколькими источниками. Так, скажем, многочисленные ленты новостей в Web сообщают об одних и тех же событиях, и в этом случае мог бы оказаться полезен инструмент, способный выделить общие места и новую информацию.

Анализ современных методов реферирования и аннотирования. В настоящее время наиболее распространено ручное аннотирование, к достоинствам которого можно отнести, безусловно, высокое качество составления аннотации – ее "осмысленность". Типичные недостатки ручной системы аннотирования – высокие материальные затраты и присущая ей низкая скорость.

Хорошее аннотирование предполагает содержание в аннотации предложений, представляющих максимальное количество тем, представленных в документе, при минимальной избыточности.

Реферирование основывается на двух подходах: общий – при создании реферата программа основывается на общих положениях создания текста (преимущество: рефераты одинаково хороши для любых тем);

специфический – при создании реферата программа уже настроена на определённые типы рефератов (например, научный, экономика).

Согласно статье [1], процесс аннотирования состоит из трех этапов: анализ исходного текста, определение его характерных фрагментов, формирование соответствующего вывода (рис. 1).

Большинство современных работ концентрируются вокруг разработанной технологии реферирования одного документа.

Выделяют два основных подхода к автоматическому аннотированию текстовых документов [4]:

Извлечение – предполагает выделение наиболее важных фрагментов (чаще всего это предложения) из исходного текста и соединение их в аннотацию (методы, основанные на этом подходе, еще называю поверхностными).

Обобщение – предполагает использование предварительно разработанных грамматик естественных языков, тезаурусов, онтологические справочники и др., на основании которых выполняется переформулирование исходного текста и его обобщение (методы, основанные на этом подходе, еще называют глубинными [4, 5]).

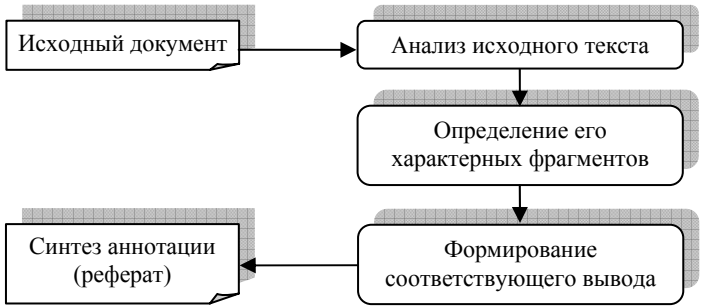

Рис. 1. Процесс аннотирования (реферирования) документа

В подходе, основанном на извлечении фрагментов методом сопоставления шаблонов, выделяют наиболее лексически и статистически значимые части (модель линейных весовых коэффициентов [3], TRM – Text Relationship Map [4]). В результате аннотация в данном случае создается простым соединением выбранных фрагментов. Этот подход легко настраивается для обработки больших объемов информации. Из-за того, что работа таких методов основана на выборке отдельных фрагментов, предложений или фраз, текст аннотации, как правило, лишен связности. С другой стороны, такой подход выдает более сложные аннотации, которые нередко содержат информацию, дополняющую исходный текст.

На рис. 2 изображена обобщенная архитектура системы автоматического реферирования, основанная на таком подходе.

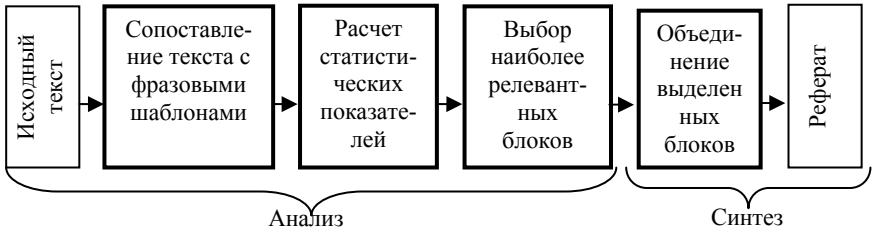

Рис. 2. Обобщенная архитектура системы автоматического реферирования

В подходе обобщения для подготовки аннотации требуются мощные вычислительные ресурсы для систем обработки естественных языков (NLP – Natural Language Processing), в том числе грамматики и словари (тезаурусы) для синтаксического разбора и генерации естественно-языковых конструкций, онтологические справочники [6].

Данный подход предполагает использование двух основных типов методов. Первый тип опирается на традиционный лингвистический метод синтаксического разбора предложения. Второй тип методов аннотирования опирается на понимание естественного языка. Синтаксический разбор также входит составной частью в такие методы анализа.

На рис. 3 изображена обобщенная архитектура системы автоматического реферирования, основанная на знаниях.

Подход, основанный на обобщении и предполагающий опору на знания, как правило, требует полноценных источников знаний. Это является серьезным препятствием для его широкого распространения. Поэтому разработчики средств автоматического аннотирования все больше склоняются к гибридным системам.

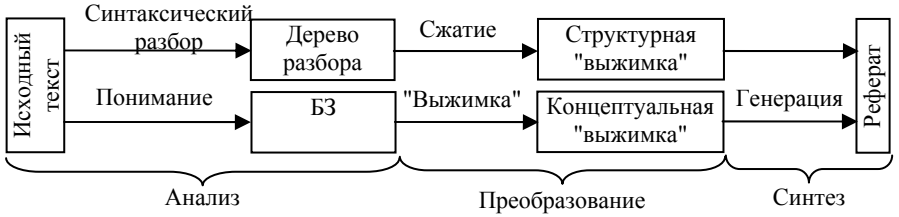

Рис. 3. Два основных подхода к формированию реферата в системах с опорой на знания

Главным ограничением обоих методов является требование сжатия. Объем аннотации, или реферата должен составлять от 5 до 30% исходного текста. Подготовка аннотаций нескольких источников информации или формирование сводок для карманных устройств предполагает еще большую степень сжатия. Добиться выполнения таких жестких требований очень

сложно, поскольку для этого необходим немалый запас знаний. Оба подхода дают примерно одинаковый результат. При их использовании раздельно либо в различных комбинациях точность лексико-грамматического анализа, скажем, английского языка улучшается до 96-98%, что сравнимо с точностью ручной обработки.

Еще одну сложность представляет оценка средств реферирования [1, 2]. Необходима гарантия того, что аннотация действительно является адекватной заменой текста, иными словами, пользователь должен быть уверен, что в кратком изложении выражены все основные мысли оригинала. Поэтому методы создания и оценки рефератов должны развиваться параллельно.

Постановка задачи и цель работы. Целью данного исследования является разработка модели синтаксического анализа предложения, охватывающие процессы реферирования и аннотирования. Синтаксический анализ и построение дерева синтаксического разбора будет реализовано с помощью алгебры предикатов и алгебры предикатных операций. Целью работы является объединить статистические методы и методы, основанные на знаниях в задачах аннотирования и реферирования полнотекстовых документов.

Моделирование синтаксического разбора в задачах аннотирования и реферирования полнотекстовых документов. Важной проблемой, возникающей при синтезе аннотаций, является отсутствие средств синтаксического и семантического анализов, а также синтеза текста на русском или украинском языках, поэтому сервисы аннотирования ориентированы либо на узкую предметную область, либо требуют участия человека.

Для частичного решения этой проблемы автором статьи предлагается использовать единый универсальный, хорошо разработанный математический аппарат. Желательно, чтобы этот математический аппарат был ориентирован также и на моделирование всех уровней лингвистической обработки текстов документов. Опыт исследования закономерностей передачи информации на естественном языке, показывает, что рационально пользоваться одним формальным аппаратом описания закономерностей передачи и интеллектуального преобразования информации. Необходимы формализмы для описания предикатов, которые реализуются при любом виде интеллектуальной обработки текстовой информации, для формирования уравнений, описывающих свойства этих предикатов. Таким наиболее универсальным аппаратом, служащим для описания закономерностей обработки информации на естественном языке, и является алгебра конечных предикатов [7].

Теперь приступим к формальному описанию предложения. Для того, чтобы увидеть в предложении формулу алгебры предикатных операций, сначала представим в виде граф-схемы синтаксическую структуру какой-нибудь формулы [8].

Возьмем, к примеру, формулу алгебры булевых функций $\overline{X_1}X_2 \vee X_3\overline{X_4}$. Ее можно выразить графически схемой, изображенной на рис. 4. Кружки со знаками булевых операций $\neg$, $\vee$ и $\wedge$ изображают преобразователи формул.

Схема синтезирует формулу $\overline{X}_1X_2 \vee X_3\overline{X}_4$ из ее аргументов X_1, X_2 X_3, X_4. Так, проходя через крайний справа блок дизъюнкции, формулы $\overline{X}_1X_2$ и $X_3\overline{X}_4$ преобразуются в формулу $\overline{X}_1X_2 \vee X_3\overline{X}_4$. Та часть формулы, на которую бинарная операция ($\vee$ или $\wedge$) действует первой, поступает на преобразующий блок по горизонтальному входу, второй – по вертикальному. Схема формулы представляет собой древовидный граф.

Древовидный граф – это частный случай параллельно-последовательной схемы. Любая формула может быть представлена в виде последовательно-параллельной схемы.

Попытаемся подойти к разработке метода построения подобных графов для предложений естественного языка. В грамматике для наглядного представления структуры предложений используются деревья синтаксического подчинения [9]. Их мы и примем в качестве отправного пункта при решении поставленной задачи.

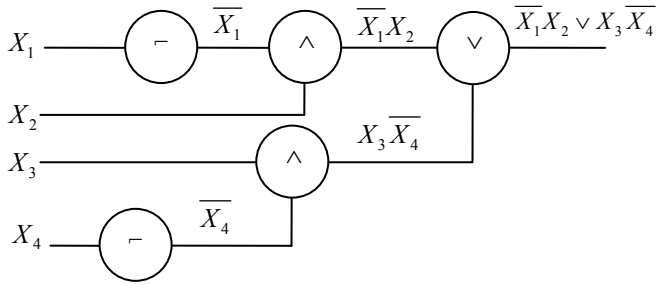

Рис. 4. Графическое представление формулы алгебры булевых функций

Пример дерева синтаксического подчинения изображен на рис. 5. Слова предложения соединяются в пары стрелками, называемыми дугами.

Рис. 5. Пример дерева синтаксического подчинения

Результат такого соединения называется словосочетанием. Слово, из которого дуга исходит, называется главным, а слово, в которое она входит – зависимым. Корнем предложения называется слово, в которое не входит ни одна из дуг.

Построение деревьев синтаксического подчинения для большого числа предложений показало, что связи между словами всегда образуют древовидную структуру, аналогичную той, которая изображена на рис. 5. Этот факт свидетельствует о том, что в предложении, кроме линейного порядка слов, существуют еще и направленные связи между словами.

Ниже описывается метод построения схемы формулы предложения. В качестве примера, возьмем предложение "Методы и алгоритмы Data Mining наиболее эффективны при анализе больших объемов данных" (Пример предложения взят из книги Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Технология анализа данных: Data Mining, Visual Mining, Text Mining, OLAP). Переходя от схемы к формуле, получаем следующую формулу алгебры предикатов, выражающую смысл предложения:

((Методы1(алгоритмы))2(Data Mining))9((наиболее3(эффективны)8((при4(анализе))7((больших5(объемов))6(данных))). (4)

Заменяя в этом выражении все слова соответствующими им предикатными переменными $X_1 \div X_{10}$, приходим к формуле алгебры предикатных операций, выражающей синтаксическую структуру рассматриваемого предложения.

$$((X_1 1 X_2) 2 X_3) 9((X_4 3 X_5) 8((X_6 4 X_7) 7((X_8 5 X_9) 6 X_{10}))), \quad (5)$$

Схема формулы предложения построена по его дереву синтаксического подчинения, так что, при желании, всегда можно возвратиться от схемы к дереву. Однако, схема содержит в себе и нечто новое, а именно: блоки, синтезирующие текст предложения из его отдельных элементов; полюсы, на которых появляются предложения и словосочетания; очередность выполнения синтеза формулы предложения блоками схемы (рис. 6). По схеме можно построить формулу предложения.

Номера выполняют в формуле роль имен операций, скобки указывают очередность их выполнения и последовательность применения каждой операции к словам, а формы слов представляют собой значения аргументов формулы.

В теории естественного языка принято, что отдельные слова выражают предикаты. Отсюда следует, что символы $X_1 \div X_{10}$ выражают предикатные переменные, номера $1 \div 10$ – предикатные операции, а само выражение (5) – формулу алгебры предикатных операций. Формула (4) выражает имя предиката предложения.

После дополнения предложения предметными переменными по методике, описанной выше, оно превращается в формулу алгебры предикатов. Итак, мы видим, что естественный язык имеет двухъярусное строение. Первый ярус представлен некоторой алгеброй предикатов, второй – алгеброй предикатных операций.

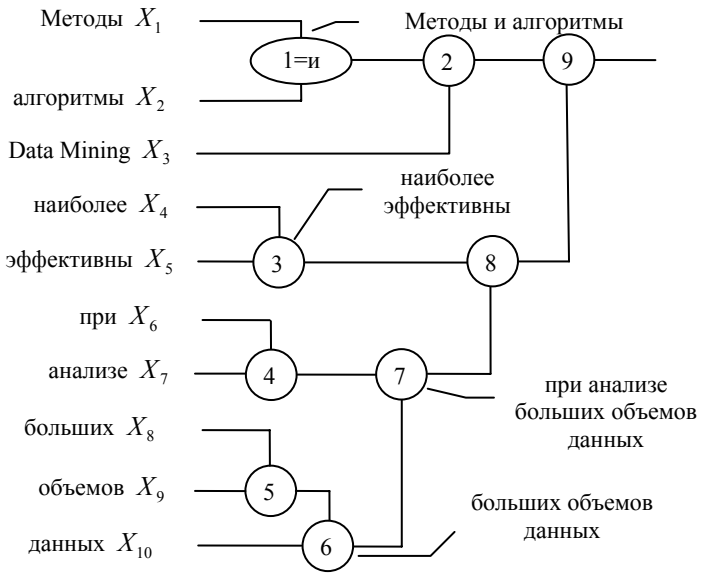

Рис. 6. Схема формулы предложения

Семантика предложения, формально описывается на языке алгебры предикатов, синтаксис, т.е. строение предложения, – на языке алгебры предикатных операций. Формула (5) показывает, в какой последовательности и из каких слов образуется предложение типа "Методы и алгоритмы Data Mining наиболее эффективны при анализе больших объемов данных".

Только что при рассмотрении синтаксической структуры предложения нам пришлось обратиться к понятию алгебры предикатных операций [10, 11, 12]. Дадим его формальное определение. Пусть U – универсум предметов; $x_1, x_2, \ldots, x_m$ – предметные переменные; P – множество всех предикатов $P(x_1, x_2, \ldots, x_m)$ на предметном пространстве U^m. Множество P называется универсумом предикатов. Переменные $X_1, X_2, \ldots, X_k$, определенные на множестве P, называются предикатными. Их значениями служат предикаты, заданные на U^m. Множество P^k называется предикатным пространством размерности k над предметным пространством U^m. Элементы множества P^k (k-компонентные наборы предикатов) называются предикатными векторами. Предикатное пространство представляет собой двухэтажную конструкцию: на ее первом этаже находятся предметы, на втором – предикаты. Любая функция $F(X_1, X_2, \ldots, X_k) = Y$, отображающая множество P^k в множество P, называется предикатной операцией. Образуем

множество R всех предикатных операций. Алгеброй предикатных операций над R называется любая алгебра, заданная на носителе R.

Пусть $F(X_1, X_2, \ldots, X_k) = Y$ – предикатная операция, отображающая множество P^k в множество P. Здесь $X_1, X_2, \ldots, X_k$ – предикатные переменные, выступающие в роли аргументов операции F; Y – предикатная переменная, являющаяся значением операции F. Отрицанием $\neg F = \overline{F}$ предикатной операции F называется такая предикатная операция, значения которой определяются по правилу

$$(\neg F)(X_1, X_2, \ldots, X_k) = \neg F(X_1, X_2, \ldots, X_k), \qquad (6)$$

для любых $X_1, X_2, \ldots, X_k \in P$. Пусть F и G – предикатные операции, отображающие P^k в P. Дизъюнкцией $F \vee G$ предикатных операций F и G называется предикатная операция, значения которой определяются по правилу

$$(F \vee G)(X_1, X_2, \ldots, X_k) = F(X_1, X_2, \ldots, X_k) \vee G(X_1, X_2, \ldots, X_k), \qquad (7)$$

для любых $X_1, X_2, \ldots, X_k \in M$. Конъюнкцией $F \wedge G$ предикатных операций F и G называется предикатная операция, значения которой определяются по правилу

$$(F \wedge G)(X_1, X_2, \ldots, X_k) = F(X_1, X_2, \ldots, X_k) \wedge G(X_1, X_2, \ldots, X_k), \qquad (8)$$

для любых $X_1, X_2, \ldots, X_k \in M$. В последних трех равенствах слева от знака равенства фигурируют операции $\neg$, $\vee$ и $\wedge$ над предикатными операциями; справа знаки $\neg$, $\vee$ и $\wedge$ обозначают операции над предикатами. Булевой алгеброй предикатных операций называется любая алгебра предикатных операций с базисом операций, состоящим из отрицания, конъюнкции и дизъюнкции.

Основные результаты и выводы. В отличие от других, ранее применяемых методов синтаксического анализа, предложенный логико-семантический метод опирается на структуру предложений и семантику текста в целом. Он позволит качественно аннотировать (реферировать) полнотекстовые документы. Качество аннотирования (реферирования) обеспечивается за счет синтаксического анализа, реализованного с помощью алгебры конечных предикатов и предикатных операций.

Список литературы: ***1.*** *Удо Хан, Индерджиет Мани*. Системы автоматического реферирования. "Открытые системы", 2000, № 12. ***2.*** *Hahn U., Mani I.* The challenges of automatic summarization. IEEE Computer, 33(11):29-35, 2000. ***3.*** *Башмаков А. И., Башмаков И. А.* Интеллектуальные информационные технологии: Учеб. Пособие. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2005. – 304 с. ***4.*** *Барсегян А. А., Куприянов М. С., Степаненко В. В., Холод И. И.* Технология анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. – СПб.: БХВ-Петербург, 2007. – 384 с.

5. *Браславский П., Колычев И.* Автоматическое реферирование веб-документов с учетом запроса. Грант ООО "Яндекс" № 102707, company/yandex. ru/grant/2005/ll_Braslavski_102707.pdf. ***6.*** *Марчук Ю. Н.* Компьютерная лингвистика: учебное пособие /Ю.Н. Марчук. – М.:АСТ: Восток – Запад, 2007. – 317 с. ***7.*** *Дударь З. В., Рассадникова А. В., Шабанов-Кушнаренко Ю. П.* Тексты естественного языка как формулы лингвистической алгебры //АСУ и приборы автоматики. – 1998. – № 107. – С. 135-144. ***8.*** *Баталин А. В.* Формальное описание структуры естественного языка как алгебры предикатных операций и его применение в системах искусственного интеллекта. – Дис. … канд. Техн. Наук. – Харьков: ХНУРЭ, 2004. – 168 с. ***9.*** *Хайрова Н. Ф., Замаруева І. В.* Машинний переклад: Навч. посіб. – Харків: Око, 1998. – 82 с. ***10.*** *Бондаренко М. Ф., Шабанов-Кушнаренко Ю. П.* Теория интеллекта: Учебник. – Харьков: ООО «Компания СМИТ», 2006. – 267-281. ***11.*** *Шабанов – Кушнаренко Ю. П.* Теория интеллекта: Проблемы и перспективы – Х.: Вища шк., 1987. ***12.*** *Шабанов-Кушнаренко Ю. П., Шаронова Н. В.* Компараторная идентификация лингвистических объектов – К., ИСИО, 1993.

Поступила в редколлегию 20.01.09