УДК 004.9

З.В. Дударь[1], С.В. Егоров[2]
[1] ХНУРЭ, г. Харьков, Украина,
[2] ХНУРЭ, г. Харьков, Украина, stas.iegorov@gmail.com

СЕМАНТИЧЕСКОЕ АННОТИРОВАНИЕ В ИНФОРМАЦИОННО-ПОИСКОВЫХ СИСТЕМАХ

Проведен анализ структуры информационно-поисковых систем, в основу работы которых положена концепция инвертированного индекса. Рассмотрены наиболее эффективные методы аннотирования текста. Предложен метод поэтапного информационного поиска, позволяющий расширить возможности пользователя и сократить время, затрачиваемое им на поиск необходимой информации.

АННОТАЦИЯ, ЗАПРОС, ИНФОРМАЦИОННО-ПОИСКОВЫЕ СИСТЕМЫ, СЕМАНТИЧЕСКОЕ СЖАТИЕ

Введение

Стремительное развитие Интернет влечет за собой развитие различных поисковых систем, сервисов и служб.

В рамках информационно-поисковых систем постоянно появляются новые сервисы, расширяется круг возможностей пользователя.

Анализ работы информационно-поисковых систем, принцип действия которых основывается на построении инвертированного индекса, позволяет расширить круг возможностей пользователя посредством предварительного ознакомления его с кратким содержанием документов, предоставляемых системой в ответ на запрос пользователя.

Необходимая реорганизация предполагает изменения в работе поискового сервера и одной из программ модуля индексирования.

Предлагаемый в работе метод поэтапного поиска информации может быть востребован в информационно-поисковых системах и позволит сократить время, затрачиваемое пользователем на поиск интересующей информации.

1. Анализ проблемы

Современные информационно-поисковые системы используют метод, в котором по запросу пользователя, состоящего из набора ключевых слов, выдается страница результатов поиска (Search Engine Result Page - SERP), содержащая ссылки на документы, находящиеся в Сети. Список составлен в порядке релевантности документов поисковому запросу. Расположение документов на странице соответствует их индексу цитирования: чем выше индекс цитирования, тем выше вероятность того, что пользователь увидит этот документ на первых страницах поисковой выдачи [1]. Далее пользователь самостоятельно осуществляет поиск нужной ему информации, просматривая каждый документ в отдельности. Пользователь вынужден тратить время на просмотр документа целиком и только по результатам просмотра судить – содержит ли документ нужную информацию или нет. Если же пользователю предлагать документ в сжатом виде, т.е. аннотацию, по которой он будет судить о присутствии искомой информации в документе, и только при наличии таковой просматривать при необходимости весь документ целиком, то временные затраты существенно сократятся.

2. Постановка задачи исследования

Целью исследования является расширение возможностей пользователя в области информационного поиска, а именно: по запросу пользователя предоставить ему возможность получения аннотации, которая полностью отражает смысл исходного текста, после прочтения которой пользователь может судить, содержит ли документ необходимую информацию. В случае позитивной оценки предоставить пользователю возможность, в зависимости от типа документа, либо загрузить полную его версию на компьютер, либо перейти на интересующий ресурс. Таким образом, предлагается использовать семантическое аннотирование текста для предоставления в распоряжение пользователя аннотации и ссылок на ресурсы сети, а в случае необходимости – предоставить полный текст документа. Таким образом, можно существенно снизить время, затрачиваемое пользователем на поиск необходимой информации.

3. Реализация

Следует отметить два аспекта: первый случай – когда пользователем корректно составлен поисковый запрос, то, получая аннотацию, он может самостоятельно принимать решение о целесообразности дальнейшего ознакомления с полным текстом исходного документа; второй случай – когда пользователь допустил неточности в формулировке поискового запроса, тогда краткие аннотации позволят гораздо быстрее обнаружить неточности в запросе и скорректировать его.

Рассмотрим последовательно техническую реализацию.

Принцип работы современных информационно-поисковых систем основан на концепции инвертированного индекса, который ставит в соответствие терминам те части документа, в которых они встречаются. Процесс создания инвертированного

индекса включает ряд этапов, важнейшими из которых являются: создание списка нормализованных лексем для каждого документа; сортировка списка, в результате которой термины располагаются в алфавитном порядке; группировка многократно повторяющихся терминов. В результате индекс представляет собой словарь, который хранит термины, расположенные в алфавитном порядке; документную частоту, равную длине списка словопозиций (принадлежность термина документу и координата термина в данном документе), а также номера документов, в которых встречается данный термин. Считается, что такая структура инвертированного индекса является достаточно эффективной для поиска текстовой информации по произвольному запросу. Однако организация инвертированного индекса по аннотациям документов, которая предлагается в данной работе, позволит существенно сократить объем обрабатываемой информации и повысить скорость работы информационно-поисковой системы.

Как известно, основными компонентами информационно-поисковых систем являются: модуль индексирования, база данных и поисковый сервер (рис. 1).

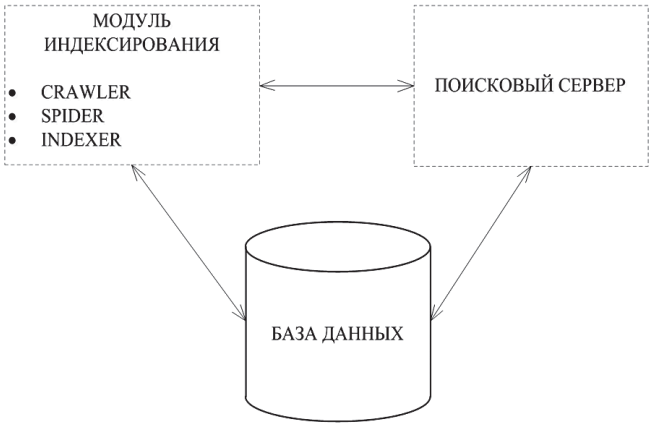

Рис. 1. Структура информационно-поисковой системы

Модуль индексирования, в свою очередь, состоит из трех вспомогательных программ:

– Spider – программа, которая осуществляет скачивание Web-страниц. Скачивается HTML-код каждой страницы;

– Crawler – программа, выделяющая ссылки, присутствующие на странице, и определяющая направление дальнейшего поиска, т.е. осуществляющая поиск новых документов, которых еще нет в системе;

– Indexer – программа, анализирующая скачанные роботами Web-страницы. Прежде всего, эта программа очищает текст индексируемой страницы от всех нетекстовых элементов, таких как: графика, теги разметки HTML. Затем осуществляется просмотр страниц с целью выборки слов, и удаления всей информации, которая словами не считается (пробелы, знаки препинания и др.). Нужно сказать, что у каждой информационно-поисковой системы есть свое определение слова в тексте. Каждая поисковая машина использует свой алгоритм лингвистической обработки слов, то есть, приведения слов к их начальным грамматическим формам или основам. Этот алгоритм называют машинной морфологией и используют для экономии места в инвертированном индексе. На следующем этапе осуществляется формирование инвертированного индекса, в котором собраны основы всех слов, а также информация об их местоположении.

Согласно предлагаемому методу, необходимо к функциям Indexer'a, таким как: анализ текста, заголовка, специальных HTML-тегов, добавить функцию аннотирования документа (рис. 2).

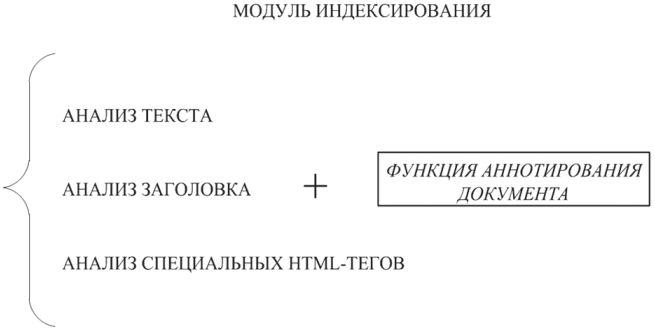

Рис. 2. Функции обновленного модуля индексирования

Алгоритм работы поискового сервера претерпевает незначительные изменения, а именно: пользовательский запрос подвергается морфологическому анализу, который передается в качестве входных параметров модулю ранжирования. Далее генерируется сниппет, в котором предлагается наряду с обычным заголовком двумя-тремя предложениями, содержащими ключевые слова, и ссылкой на сам документ дать пользователю возможность при наведении курсора на ссылку, кроме просмотра сохраненной в кэше копии документа или похожих документов, ознакомиться и с аннотацией.

Таким образом, пользователь получает возможность уже на начальном этапе поиска иметь представление о документах, содержащих искомую информацию, и в случае необходимости просмотреть тексты документов.

Такая техническая реализация позволит также существенно сократить время, затрачиваемое системой на поиск нужной информации.

Важным является тот факт, что предоставляемая пользователю информация должна полностью отражать смысл исходного текста.

На данный момент существует несколько различных подходов к составлению аннотаций: семантическое сжатие, сжатие текста путем замещения терминов их ближайшими аналогами, а также

так называемое «умное кодирование», в основе которого лежит совмещение различных подходов к кодированию разнородной информации. Например, использование различных методик для кодирования текста, чисел, IP и URL-адресов в рамках одного документа.

Поскольку аннотирование само по себе является достаточно сложной задачей, то, как правило, собственно аннотированию предшествуют подготовительные этапы. Одним из них может являться упрощение исходного текста, которое заключается в формировании из синтаксически сложных предложений исходного текста, содержащих различные грамматические обороты, «легкодоступных предложений» (Easy Access Sentences - EAS). Такие предложения содержат только один глагол и имеют максимально простую синтаксическую структуру.

Создание таких «легкодоступных предложений» необходимо с целью упрощения анализа текста системами машинного перевода, извлечения информации, а также – системами аннотирования текстов.

Легкодоступные предложения – это попытка объединения насыщенных информацией естественноязыковых данных с приложениями, разработанными для эффективного использования структурированных, хорошо организованных данных, которые, однако, трудно получить без существенного человеческого вмешательства [2].

Самым сложным при формировании EAS является точная передача смысла исходного текста. EAS представляют собой предложения, сформированные путем декомпозиции исходных предложений на более простые части. Поэтому формирование EAS включает в себя такие этапы, как: выявление имен, установление семантической зависимости между именами и соответствующими частями предложения, выделение этих частей из текста, нахождение глаголов и приведение их в соответствующую временную форму, формирование простого утвердительного предложения с использованием имени, глагола и оставшейся части предложения.

Теоретической основой сжатия является такое свойство текстовых сообщений, как повторы частей информации в текстах. Это свойство широко используется в информационных системах при создании поискового образа документа. Частота упоминаний значимых слов говорит об их важности для содержания текста. Задача заключается в отделении значимых слов от простых. Семантическая значимость фрагмента учитывает синтаксические индикаторы, лексические повторы, а также разного рода вспомогательные списки, которые дифференцируют лексику текста. Обычно семантический компонент работает «снизу» с интерпретации синтаксического представления. Система смысловой компрессии интерпретирует связи между словами и целыми предложениями таким образом, чтобы стало меньше узлов и связей в структуре, но их содержательный вес повысился. Для этой цели разработана словарная классификация лексем по степени информативности. Таким образом, в сложный узел попадут как имя объекта со всеми его атрибутами, так и имя действия с обозначениями модальности, стадии и времени.

На следующем этапе «снизу» осуществляется разбиение текста на другие фразы или простые предложения и построение других объединений, которые являются продолжением или дополнением предиката с большим весом. Таким образом, в тексте намечаются основы ситуаций, а дальнейшее их оформление требует выхода за пределы текстового материала [3].

Поскольку речь идет об аннотировании текста, то особый интерес представляют методы необратимого сжатия, то есть сжатия информации с потерями. Известно, что алгоритмы сжатия информации с потерями делятся на семантически независимые и семантически зависимые. Алгоритмы первого типа в свою очередь представлены адаптивными и статистическими алгоритмами. Адаптивные алгоритмы осуществляют сжатие текста путем его однократного сканирования. Кодирование в этом случае происходит посредством нахождения повторяющихся участков текста и замене их указателями, адресованными к фрагменту текста, где такой участок уже встречался. Следует отметить, что адаптивные алгоритмы сравнительно медленны. Время кодирования у них непосредственно зависит от длины исходного текста, а достижимое сжатие определяется машинной реализацией алгоритма.

Статистические алгоритмы используют словари, представленные либо кодовой таблицей символов алфавита, либо словарем фрагментов переменной длины. В зависимости от избранного вида словаря различают и способы реализации алгоритмов: кодирование фрагментов фиксированной или переменной длины.

Информационно-поисковые системы используют, как правило, способ кодирования фрагментов кодами фиксированной длины, в которых используются двухбайтовые коды. Однако такое кодирование дает небольшой коэффициент сжатия.

Наиболее эффективным методом кодирования фрагментов кодами переменной длины является метод Хаффмана, который для идентификации использует поиск по двоичному дереву [4]. Следует заметить, что коды переменной длины используют для словарей больших размеров, содержащих как буквы алфавита, так и фрагменты текста.

Особый интерес для целей информационного поиска представляют семантически зависимые методы, опирающиеся на грамматику естественного языка. При формировании аннотаций используют

принципы автоматического обобщения текста [5]. Обобщение может представлять уменьшающее преобразование исходного текста путем сжатия содержания с помощью выделения и/или обобщения той информации, которая является важной [6]. Наиболее эффективно использование принципа автоматического обобщения текста для обработки больших объемов документов по схожей тематике, а также в области обработки естественного языка (Data Mining и понимание текста).

Перспективным с точки зрения семантического сжатия текста представляется новый подход, основанный на «принципе количества кода». В основу принципа положена гипотеза о том, что большее количество информации будет закодировано большим количеством кода.

Основной проблемой автоматического обобщения текста является его количественная оценка, которая предполагает сопоставление созданных автоматически обобщений с моделями, разработанными человеком. Одними из последних разработок в области автоматических методологий оценок обобщений является метрика ROUGE и метод Пирамиды.

Обработанный таким образом исходный текст представляет собой набор синтаксически простых предложений, легко поддающихся анализу системами аннотирования текстов.

Выводы

Проведенное исследование основных принципов работы информационно-поисковых систем позволило внести коррективы в процесс формирования страницы результатов поиска в ответ на запрос пользователя.

В работе предложен метод поэтапного информационного поиска, который поможет расширить возможности пользователя при нахождении нужной информации и существенно сократить время, затрачиваемое на поиск.

Метод предполагает, что при работе пользователя в информационно-поисковой системе в ответ на его запрос необходимо предоставлять аннотацию, полностью отражающую смысл исходного текста, и ссылки на местоположение исходного документа. По результатам обзора аннотации пользователь может самостоятельно принимать решение о необходимости ознакомления с полным текстом документа в случае обнаружения в аннотации интересующей информации.

Предложенный метод предусматривает внесение корректив в работу модуля индексирования. Так, к обычным функциям анализа текста, заголовка и специальных HTML-тегов следует добавить функцию аннотирования документа.

Проведенный в работе анализ методов аннотирования показывает, что для создания аннотаций в методе поэтапного информационного поиска наиболее эффективными являются семантически-зависимые методы сжатия. Особый интерес представляют методы, основанные на принципах автоматического обобщения текста и использующие новый подход «принципа количества кода».

Предложенный в работе метод поэтапного поиска предназначен для использования в информационно-поисковых системах и позволяет качественно улучшить условия работы пользователя в системе, существенно повысить скорость обработки запроса и предоставить новые возможности в области информационного поиска.

Список литературы. 1. *Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце*. Введение в информационный поиск : Пер. с англ. — М. : ООО «И.Д. Вильямс», 2011. – 528 с. **2.** *Klebanov B. B., Knight K., Marcu D.* Text Simplification for Information-Seeking Applications. – Springer Verlag, 2004. – 13 p. **3.** *Леонтьева Н.Н.* Автоматическое понимание текстов: системы, модели, ресурсы //учеб. пособие для студ. лингв, фак. вузов. — М.: Издательский центр «Академия», 2006. – 304 с. **4.** *Alonso i Alemany, L., Fuentes Fort, M.*: Integrating cohesion and coherence for automatic summarization. In: EACL '03: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics. (2003) 1–8. **5.** *Lloret, E., Ferrández, O., Miñoz, R., Palomar, M.*: A Text Summarization Approach Under the Influence of Textual Entailment. In: Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2008) 12-16 June, Barcelona, Spain. (2008) 22–31. **6.** *Дударь З.В., Егоров С.В.* Исследование и оптимизация методов сжатия текстовой информации // Вестник ХНТУ. – Херсон. – 2012. – № 1(44). – 65-68 с.

Поступила в редколлегию 04.02.2013

УДК 004.9

Семантичне анотування в інформаційно-пошукових системах / З.В. Дудар, С.В. Єгоров // Біоніка інтелекту: наук.-техн. журнал. – 2013. – № 1 (80). – С. 104-107.

У роботі запропоновано метод поетапного інформаційного пошуку, що допоможе розширити можливості користувача при знаходженні потрібної інформації та істотно скоротити час, що витрачається на пошук.

Метод передбачає, що користувач у відповідь на свій запит отримує анотацію та посилання на місцезнаходження вихідного документу. За результатами огляду анотації користувач може самостійно приймати рішення щодо необхідності ознайомлення із повним текстом документу.

Іл. 2. Бібліогр.: 6 найм.

UDC 004.9

Semantic annotation at information storage and retrieval system / Z. Dudar, S. Iegorov // Bionics of Intelligense: Sci. Mag. – 2013. – № 1 (80). – P. 104-107.

Method of step-by-step information retrieval which will help to enlarge user possibilities while searching for necessary information and dramatically decrease search time was proposed in the work.

Method provides for giving an annotation and reference to location of source text per user's request. By results of brief review of the annotation user can make decision of necessity of full original text viewing.

Fig.2. Ref.: 6 items.