условиях отличных от рекомендуемых производителем в силу неконтролируемого человеческого фактора от 20 до 30% [10 – 12].

Список литературы: 1. ГОСТ Р 51188-98 Защита информации. Испытания программных средств на наличие компьютерных вирусов. – М., 1998. – 8 с. **2.** ГОСТ 27.301-95 Надежность в технике. Расчет надежности - основные положения. – М., 1995. – 19 с. **3.** Military Handbook “MIL-STD-810F” Environmental Engineering Considerations and Laboratory Tests, 2000. – 539 p. **4.** Military Handbook “MIL-STD-883G” Department of Defense. Test Method Standard. Microcircuits, 2006. – 716 p. **5.** IPC-SM-785 Guidelines for Accelerated Reliability Testing of Surface Mount Solder Attachments, 1992. – 58 p. **6.** *Критенко М.* Обеспечение качества военной продукции: Новое поколение нормативных документов / М. Критенко // Электроника: Наука, Технология, Бизнес. – 2000. – №4. – С. 50–53. **7.** *Лунтовський А.О.* Технології розподілених програмних додатків. Монографія – К.: Державний університет інформаційно-комунікаційних технологій "ДУІКТ", 2010. – 474 с. **8.** *Козлов Б.А.* Справочник по расчету надежности аппаратуры радиоэлектроники и автоматики / Ушаков И.А. – М.: Сов. радио, 1975. – 472 с. **9.** *Корн Г.* Справочник по математике / Корн Т. – М.: 1974. – 830с. **10.** Military Handbook “MIL-HDBK-338B” Electronic Reliability Design Handbook, 1998. – 1046 p. **11.** Military Handbook “MIL-HDBK-217F” Reliability Prediction of Electronic Equipment, 1991. – 205 p. **12.** Military Handbook “MIL-HDBK-344A” Environmental Stress Screening (ESS) of Electronic Equipment, 1993. – 102 p.

Поступило в редколлегию 15.02.2012

УДК 004.4'414

Е. А. ГРИДИНА, студ., ХНУРЭ, Харьков

АНАЛИЗ АЛГОРИТМОВ АВТОМАТИЧЕСКОГО АННОТИРОВАНИЯ ТЕКСТА НА ОСНОВАНИИ СЕМАНТИЧЕСКОГО ПРЕДСТАВЛЕНИЯ

Дана загальна характеристика методів і алгоритмів автоматичного анотування тексту, таких як базовий алгоритм на основі розбору HTML та алгоритм SRL. Описані основні кроки використання алгоритмів для досягнення оптимального результату в складанні анотації.
Ключові слова: анотація, SRL, HTML, метод, пасаж.

Дана общая характеристика методов и алгоритмов автоматического аннотирования текста, таких как базовый алгоритм на основании разбора HTML та алгоритм SRL. Описаны основные шаги применения алгоритмов для достижения оптимального результата в составлении аннотации.
Ключевые слова: аннотация, SRL, HTML, метод, пассаж.

This article represents general features of methods and algorithms of automatic annotation of the text, such as basic algorithm based on HTML extracting and SRL algorithm. Main steps of using different algorithms to achieve optimal results in the preparation of abstracts have been described.
Keywords: abstract, SRL, HTML, method, passage.

1. Введение

Искусство реферирования, или составления аннотаций, стало неотъемлемой частью повседневной жизни. Новости, которые предлагает нам телевидение, – это суть реферат мировых событий дня.

В настоящее время известно много алгоритмов автоматического аннотирования или формирования краткого содержания документов, например МЛ Аннотатор, Золотой ключик, TextAnalyst, системы IBM Intelligent Text Miner,

Oracle Context и Inxight Summarizer (компонент АИПС AltaVista), также алгоритмы аннотирования широко используются в поисковых системах, таких как Google, Яндекс, Рамблер [1].

Зачастую возможности систем ограничены выделением и выбором оригинальных фрагментов из исходного документа и соединением их в короткий текст. 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.

Для достижения наиболее оптимального результата, необходимо совместить несколько алгоритмов аннотирования. В данной работе нами будут рассмотрены алгоритма формирования контекстно-зависимых аннотаций на основании семантического представления.

2. Описание методов автоматического аннотирования

За методом построения аннотации алгоритмы можно разделить на две группы: генерирующие и вытягивающие [2].

Генерирующие алгоритмы анализируют выходной документ или документы, связанные с ним, для поиска информации, на основе которой генерируют текст аннотации. В отличии от генерирующих вытягивающие алгоритмы аннотирования формируют аннотацию, используя текстовые фрагменты документа или его контекста. Стоит отметить, что много подходов представляют собой смешанные алгоритмы, которые используют различные методы обработки информации для построения аннотации.

Рассмотрим некоторые алгоритмы для формирования контекстно-зависимых аннотаций.

Первый алгоритм – алгоритм контекстно-зависимого реферирования HTML документа [3].

Особенность данного алгоритма заключается в том, что при составлении аннотации существенную роль играют не только сами слова запроса, но и слова, очень близкие по смыслу слов запроса (назовем их слова-переходы), а также для некоторых слов запроса и транслитерации, т.е. написание этих слов латиницей. Для составления базы таких слов-переходов использовался специальный алгоритм морфологического анализа и поиска однокоренных слов и слов-синонимов.

Для составления хороших аннотаций предварительно осуществляется выделение из структурных элементов DOM-модели HTML-документа релевантных и значимых блоков текстового контента. В результате остается серия текстовых блоков без дополнительной разметки, содержание которых наиболее полно отражает основное содержание документа, по которому и строится аннотация. Алгоритм выделения значимых блоков следующий.

Шаг 1. Обработчику (программе, реализующей выполнения алгоритма) указывается url адреса рассматриваемой страницы, по указанному адресу производится загрузка HTML-документа.

Шаг 2. Преобразование HTML в XML.

Шаг 3. Предыдущая редукция структуры XML-документа. Этап включает: удаление всех не значащих текстовых и блочных элементов, преобразование иерархической структуры документа заменой каждого из ее узлов контент-узлом.

Шаг 4. Редукция структуры документа индикаторами

Шаг 5. Извлечение текста из документа. На пятом шаге работы алгоритма содержимое корневого контент-сайта полностью сокращенной модели документа выводится в текстовый файл.

После выделения значимых и релевантных структурных элементов документа проводится аннотирование документа. Алгоритм составления аннотации включает в себя следующие шаги:

1. Получение контента значимых и релевантных элементов документа.

2. Разбивка контента на пассажи. Разбивка текста на пассажи происходит знаками препинания, обозначающий конец предложения, а также по так называемым "разделителем" html-тегов, таких как <p> </ br>, , <li>, <td> и некоторым другими.

3. Формирование кластеров с N пассажей.

4. Расчет параметров ранжирования пассажей и взвешенной оценки ранга каждого кластера.

5. Ранжирование кластеров по взвешенной оценки и определения кластера пассажей с наибольшим рангом.

6. Запуск алгоритма обрезки аннотации к $n \leq k$ символов (в данном случае $k = 300$) и формирование аннотации.

Далее срабатывает стандартный алгоритм обрезки аннотации по максимально возможному количеству слов [4].

Этот тип алгоритма считается наиболее легким, так как не учитывает многих моментов формирования связных аннотаций. Особого внимания в данной ситуации заслуживает алгоритм на основании семантической маркировки (SRL) [5].

В основу алгоритма входит блок анализа семантических структур аргумент-предикатов. На основании концепции описания языка для естественных языков (CDL.nl), предназначенная для описания концепции текста с помощью набора заранее определенных семантических отношений. Предполагается, что все связи в тексте заранее известны. Происходят семантические маркировки связей, после чего эту маркировку проверяет специальная предикат-структура, целью которой является обеспечение связности предложений в тексте.

Внутренние связи: связи подразделяются на роли, разделение на 6 абстрактиних связей: QuasiAgent, QuasiObject, QuasiInstrument, QuasiPlace, QuasiState и QuasiTime. Кроме того, каждая абстрактная связь включает в себя несколько конкретных отношений, которые выражают конкретную семантическую информацию. Например, QuasiAgent содержит пять конкретных семантических связей: agt (агент), aoj (атрибуты), cag (под-агент), cao (составляющая с атрибутами), ptn (партнер).

Квалификация отношений: в дополнение к каждому конкретному случаю, относящийся к типам, CDL.nl также определяет квалификацию связей типов. Есть девять квалификаций связей, в целом содержат mod (модификации), pos

(процессор) и qua (количество). Это подмножество связей имеет важное значение для описания предложения с множеством свойств.

Метод Кернела для классификации связей показывает связи каждой пары элементов со специфическим набором CDL.nl связей, а также описывает классификацию связей, которая использует кернелевские функции для моделирования знаний о 3-х уровнях обработки речи: синтаксический анализ, дерево зависимостей и лексическая конструкция.

Синтаксис описывает как общие так и синтаксические функции слов. Например, % NH (номинальный узел) и%> N (определитель от номинального), общие синтаксические теги @ SUB (Subject) и @ F-Subj (формальные темы) являются семантическими тегами функции. Парсер определяет 40 тегов синтаксиса.

Для каждой пары экземпляров строим дерево зависимости - извлекаем набор зависимостей путем определения токена зависимости DT = (DEP, PATH), где DEP содержит две метки: первая метка регулируется непосредственно заглавным словом главного набора, а вторая указывает на главное слово дочерней зависимости.

Таким образом, аннотация формируется с помощью трехуровневой обработки, что дает возможность формировать связную и грамотную аннотацию. Этот метод гораздо более трудоемкий по сравнению с предыдущим, но это дает свои результаты в виде хорошего реферата. Также усложненные варианты второго алгоритма активно используются в поисковых системах, чего нельзя сказать о первом алгоритме.

4. Вывод

Сравнительный анализ двух алгоритмов показал следующее - алгоритм контекстно-зависимого реферирования HTML документа более легкий и простой в применении и реализации, однако, аннотации, составлены по этому алгоритму просты, состоят из слов, которые уже есть в тексте, что значительно ограничивает возможности составления аннотации.

Алгоритм на основании семантической маркировки (SRL) изучает семантические структуры аргумент предикатов. Определяя связи в словах, их род, число, падеж и так далее, алгоритм способен заменять, выбрасывать, сокращать слова в тексте. Аннотация становится полной и законченной, а также нет ощущения повторения одного и того же в основном тексте и в аннотации.

Список литературы: 1. *A. Tombros, M.* Sanderson. Advantages of Query Biased Summa-ries, in proc. of the ACM SIGIR, 1998 **2**. *I. Mani.* Automatic Summarization, JohnBenjamin's Publishing Company, 2001. **3**. *I. Mani,* Summarization Evaluation: An Overview, in proc. of the NTCIR Workshop, 2001 **4**. *М. Губин, А. Меркулов.* Эффективный Алгоритм Формирования Контекстно-Зависимых аннотаций, Диалог 2005 **5**. Аннотрование и реферирование [Електронний ресурс] / - Режим доступу: http://www.publisher.ssu.samara.ru/archive/2003/files/20030352.pdf - 17.05.2011 г. - Загл. з екрану.

Поступила в редколлегию 15.02.2012