Автоматическое реферирование веб-документов с учетом запроса*

Павел Браславский
ИМАШ УрО РАН
pb@imach.uran.ru

Иван Колычев
УГТУ-УПИ
kis@datakrat.ru

Аннотация

В отчете описаны принципы функционирования экспериментальной системы реферирования Веб-документов. В системе используется восходящий к работам 1950-60-х годов подход к выделению важных предложений (*sentence extraction*) с небольшими модификациями. Приводятся примеры работы системы, оценка ее производительности. Описаны эксперименты по оценке качества работы системы. Предварительные результаты позволяют говорить о приемлемом качестве реферирования. Предложенный подход может быть эффективно реализован в современных информационно-поисковых системах, если поисковый индекс хранит позиции слов с привязкой к предложениям.

1. Введение

Рефераты, конспекты и другие виды сокращенных представлений текстовых документов используются со времен появления письменности. Ясно, что представление текста, передающее основное содержание и опускающее детали, может быть полезно во многих контекстах.

В обзоре [4] приводятся следующие типы рефератов:

- *индикативные* (должны предоставлять достаточно информации для принятия решения, есть ли необходимость обращаться к оригиналу);

* Работа выполнена при поддержке ООО «Яндекс» (www.yandex.ru), грант № 102707.

- *информативные* (должны заменять собой первоисточник, содержать фактическую информацию в сжатом виде);
- *критические* рефераты не только передают основное содержание документа, но и дают ему оценку.

Кроме того, рефераты могут быть *общими* или *специализированными* – соответственно, ориентированными на широкий круг читателей или на определенную группу – профессиональную, возрастную, культурную.

Интернет делает потенциально доступными огромные объемы информации и, тем самым, ставит новые проблемы – эффективной работы с такими объемами. В ситуации «информационной перегрузки» особенно актуальными становятся автоматические методы работы с большими объемами информации, в частности – методы получения сжатого представления текстовых документов – рефератов, или аннотаций.

В обзоре [4] указывается на два основных подхода к автоматическому реферированию. Первый подход ориентирован на извлечение фрагментов (обычно – предложений, отсюда общее обозначение подхода – *sentence extraction*) исходного текста, из которых и составляется реферат. Второй подход предполагает использование более изощренных методов лингвистического и семантического анализа. В данном случае обычно говорят о генерации реферата (*summary generation*) на основе семантического представления текста. Результат работы систем реферирования, основанных на извлечении предложений, далек от идеала – связного реферата, составленного квалифицированным специалистом. Однако более качественные системы реферирования требуют сложного программного обеспечения, имеют более низкую производительность и часто налагают существенные ограничения на стиль и тематику исходного текста.

Задача автоматического составления короткого реферата текстового документа возникает в контексте Веб-поиска. В списке результатов поисковой машины наряду с заголовком и адресом обычно присутствуют сниппеты (англ. *snippets*) – фрагменты документа, содержащие слова запроса. Назначение сниппетов – помочь пользователю составить представление о документе и решить, имеет ли смысл обращаться к оригиналу.

В данном отчете описана экспериментальная система автоматического реферирования Веб-документов (в том числе – с учетом запроса) *eXtragon*. В системе используется традиционный подход к выделению важных предложений (*sentence extraction*). Дополнения

метода касаются учета HTML-разметки, попыток учета не только отдельных слов, но и словосочетаний, а также введения требований к длине и разнообразию предложений реферата. Приводятся примеры работы системы, оценка ее производительности. Описаны эксперименты по оценке качества работы системы. Предварительные результаты позволяют говорить о приемлемом качестве реферирования. Предложенный подход может быть эффективно реализован в современных информационно-поисковых системах, если поисковый индекс хранит позиции слов с привязкой к предложениям.

2. Состояние исследований

Основные принципы и подходы для извлечения значимых предложений из текста на основе формальных параметров были сформулированы еще в конце 50-х – начале 60-х годов XX века [2, с. 125-129]. Эти методы включают в себя выделение значимых (ключевых) слов текста, вычисление весов предложений на основе весов входящих в него слов и дополнительных критериев: положения предложения в тексте (начало и конец документа, начало абзаца и т.д.), наличия «сигнальных фраз». В свою очередь, процедура выделения ключевых слов (или подсчета весов слов) может использовать частоту встречаемости слов в тексте и во всей коллекции, встречаемость слова в заголовках и т.д.

Такой подход к составлению рефератов широко используется до настоящего времени. Так, в работе [5] описывается система *ANES*, в которой важные слова выделяются на основе подхода TF*IDF. Пожалуй, единственное отличие от традиционного подхода – обнаружение анафорических оборотов, что позволяет исключить из реферата предложения с «висящими» отсылками к предыдущим предложениям.

В работе [12] предлагается концепция *фрактального реферирования* для поставки контента на мобильные устройства, которые обычно характеризуются малыми размерами экрана и вычислительной мощностью, низкой скоростью передачи данных. Предложенный подход использует традиционные методы выделения предложений, но дополнительно учитывается информация о иерархической структуре документа и требуемый «уровень абстракции» представления документа.

Методы *симметричного реферирования*, описанные в [3, 7], учитывают прежде всего связи между предложениями: важными считаются предложения-концентраторы связей. Связи между предложе-

ниями выявляются на основе готового словаря терминов предметной области, что несколько ограничивает применимость метода.

Подход на основе разнообразия предложен в [9]. Идея метода состоит в том, чтобы сначала найти тематические кластеры документа (т.е. группы предложений, относящихся к одной подтеме документа). После этого важные предложения выделяются из каждого кластера с использованием традиционных методик. Кластеризация предложений производится с помощью модифицированного метода *k*-средних.

Ряд работ специально посвящен проблеме составления рефератов по запросу.

В работе [10] описаны эксперименты по оценке качества и полезности рефератов, составленных с учетом запроса, в задачах информационного поиска. Метод реферирования основывается на простом подсчете частоты терминов в документе, учете терминов заголовка, терминов запроса, а также положения предложения в тексте. В эксперименте использовалась коллекция TREC (документы, запросы и оценки релевантности) и «классическая» информационно-поисковая система. В ходе эксперимента пользователи делали заключения о релевантности документов, основываясь на (1) рефератах, сформированных с учетом запроса, и (2) начальных фрагментах документа. Эксперимент показал, что использование рефератов, сформированных с учетом запроса, дает более высокие показатели точности и полноты; пользователи обрабатывают больше информации, реже обращаются к полным текстам документов.

В статье [11] описывается аналогичный эксперимент. Разработанная система автоматического реферирования *WebDocSum* использует традиционный подход извлечения предложений. Вес предложения вычисляется на основе следующих параметров: присутствие в предложении слов заголовка, слов запроса, выделенных слов (курсивом, полужирным шрифтом или подчеркиванием), а также положение предложения (в начале или конце документа, абзаца). Интересно отметить, что частотные характеристики слов в данном случае вообще не используются. В эксперименте анализировались четыре системы: *AltaVista, Google, AltaVista+WebDocSum, Google+WebDocSum* с унифицированным пользовательским интерфейсом. Участникам эксперимента были предложены четыре задания по поиску. По результатам эксперимента машины поиска в сочетании с системой реферирования были более эффективны при выполнении заданий.

В работе М. Губина и А. Меркулова [1] предлагается подход к выделению важных слов для последующего формирования сниппе-

тов на основе алгоритма *LRU-K* (вариант метода формирования буфера данных «последний недавно использовавшийся»). Такой подход позволяет учесть «локальную плотность» распределения слов в документе. Дополнительно учитывается структура документа: рефераты формируются для каждой части документа отдельно, далее отбираются пять фрагментов с наибольшим весом. Эксперименты с участием экспертов показали превосходство предложенного метода над методами, основанными на частотах слов. Кроме того, предложенный метод продемонстрировал высокую производительность.

3. Идея исследования

Эксперименты, описанные в [10, 11] демонстрируют, что даже простые методы составления рефератов с учетом запросов позволяют повысить эффективность поиска информации.

Содержание нашего исследования состоит в разработке системы реферирования Веб-документов, основанной на традиционном подходе извлечения предложений, и оценке эффективности подхода в задачах Веб-поиска.

Для вычисления весов отдельных слов мы используем подход TF*IDF, присутствие слов в заголовке и подзаголовках документа, выделение средствами HTML (обычно эта информация уже содержится в индексе машин поиска). Слова запроса получают дополнительные «очки». Кроме того, мы попытались учесть в анализе важные словосочетания, используя простой метод на основе частот пар слов и морфологических шаблонов, описанный в [8].

Вес предложения вычисляется на основе весов входящих в него слов и словосочетаний. Дополнительно увеличивается вес предложений в начале и конце документа. Вес предложения, в котором встретилось больше одного слова запроса, увеличивается нелинейно.

В ходе предварительных экспериментов мы пришли к выводу, что для бОльшей информативности и удобочитаемости реферата надо понижать вес как очень коротких (до трех слов), так и очень длинных предложений. Для экспериментов мы произвольно установили «оптимальную» длину предложения в десять слов. Кроме того, при экспериментах с Веб-документами рефераты часто состояли из очень похожих (или одинаковых) предложений. Поэтому мы ввели дополнительное требование разнообразия предложений реферата. Такое решение можно рассматривать как вариант принципа разнообразия, предложенного в [9].

Другой важной задачей исследования была оценка качества работы системы. Система *eXtragon* участвовала в дорожке по контекстно-зависимому аннотированию РОМИП-2005 (результаты должны быть известны в августе 2005 года). Кроме того, мы разработали планы двух экспериментов по сравнительной оценке качества реферирования. К сожалению, мы смогли провести эти эксперименты только с небольшим количеством участников.

4. Принципы и алгоритмы работы системы

Внешние ресурсы

Для хранения списка стоп-слов, частотного словаря и слов текста использовался метод динамического хэширования HashTrie, реализованный SoftLab MIL-TEC Ltd.[1]

Для приведения слов частотного словаря, текста и запроса к словарной форме, а также для установления грамматических характеристик словоформ для выделения словосочетаний использовался морфологический анализатор *mystem* (разработчики – В. Титов, И. Сегалович, компания «Яндекс»).[2]

Исходный частотный словарь был предоставлен компаний Яндекс и содержал 1 000 000 наиболее частых слов из русских букв в базе Яндекса. Мы обработали частотный словарь с помощью *mystem* (использовались первые варианты анализатора словарных форм, возвращаемых анализатором) и «подправили» частоты по документам (IDF) для словарных форм. В используемом нами частотном словаре содержится 333 605 слов.

Система разработана в среде Delphi.

Этапы обработки документа

Предварительная обработка. Загружается HTML-файл, кодировка меняется на Windows-1251 (поддерживаются KOI8-R, UTF-8), удаляются описания стилей, сценарии и т.д. Удаляются все HTML тэги, кроме <P>, <B>, <I>, <U>, <TITLE>, <H1>, <H2>, <H3> и <H4>. Escape-последовательности меняются на соответствующие символы. Последовательности непечатных символов (пробелы, табуляции, переводы строк) меняются на единичный пробел.

Вычисление базовых (частотных) весов слов. С помощью морфологического анализатора *mystem* все слова текста приводятся к

[1] См. http://www.softcomplete.com/products/hashtrie/hashtrie.asp
[2] См. http://corpora.narod.ru/mystem

словарной форме (если анализатор возвращает несколько вариантов, то берется первый); для прилагательных и существительных сохраняется грамматическая информация о словоформе (род, число, падеж), которая используется для нахождения важных словосочетаний. Слова из латинских букв, числа, стоп-слова не учитываются. Базовый вес слова вычисляется по формуле TF*IDF.

Вычисление базовых весов словосочетаний. В системе использован простой подход к выделению важных словосочетаний на основе подсчета частот пар слов и морфологических шаблонов [8]. Мы использовали шаблоны *A+N* (согласованное прилагательное + существительное, например: «первая леди», «последний звонок») и $N+N_G$ (существительное + существительное в родительном падеже, например: «замок зажигания», «карта мира»). Базовый вес словосочетания вычисляется как удвоенный TF*IDF менее редкого слова, входящего в словосочетание. Далее словосочетания рассматриваются наряду с другими ключевыми словами.

Вычисление весов ключевых слов. Для вычисления окончательного веса слова базовый (частотный) вес умножается на коэффициент

$K = 1 + K_B + K_U + K_I + K_T + K_H + K_Q$,

где

K_B, K_U, K_I – равны двум, если слово выделено соответственно жирным шрифтом, подчеркиванием или курсивом;

$K_T = 10$, если слово присутствует в заголовке;

$K_H = 5$, если слово встречается в подзаголовках H1..H4;

$K_Q = 500$, если слово присутствует в запросе.

Все коэффициенты подобраны эмпирически, могут настраиваться в программе.

Вычисление весов предложений. Граница предложения определяется на основе шаблонов:

- [.|?|!|…] [пробел] [A..Z|A..Я|0..9|-];
- [<P>|</TITLE>].

Итоговый вес предложения вычисляется по формуле:

$$P = L \cdot I \cdot e^{-(\frac{SL-OL}{K})^2} \cdot (1 + \frac{2q^2}{QL}) \cdot \sum_{i=1}^{SL} W_i ,$$

где

L – повышающий коэффициент для первых и последних четырех предложений документа;

I – понижающий коэффициент для вопросительных предложений;

SL – длина предложения в словах;

QL – длина запроса в словах;

W_i – вес i-го слова в предложении;

q – количество слов запроса в предложении;

OL – «оптимальная» длина предложения для реферата;

K – коэффициент уменьшения веса предложения при отклонении от «оптимальной» длины.

Формирование реферата с заданным количеством предложений. Предложения сортируются в соответствии с вычисленным весом по убыванию. Первое предложение помещается в реферат. Каждое следующее предложение берется из списка и сравнивается с предложениями реферата. Предложение отбрасывается, если оно имеет 80% или более общих слов с предложениями реферата. Процесс повторяется, пока не будет отобрано заданное количество предложений. Отобранные предложения выдаются в том порядке, в каком они находились в тексте.

Формирование реферата заданной длины (в символах). По заданию дорожки по контекстно-зависимому аннотированию в рамках РОМИП-2005 мы должны были формировать рефераты длиной до 300 символов. Для этого мы выделяли по одному фрагменту из каждого предложения реферата вплоть до исчерпывания лимита; итоговый реферат был набором таких фрагментов. Для формирования фрагмента в качестве базиса выбиралось самое важное слово в предложении (или фрагмент предложения, ограниченный словами запроса). Если лимит символов не исчерпан, стандартным контекстом отсечения считались по пять слов слева и справа от базиса плюс слова до ближайшего знака препинания или конца/начала предложения. Если базис находился на расстоянии ближе пяти слов к концу или началу предложения, то длина контекста (10 слов) перераспределялась в соответствующем направлении. При достижении лимита символов контекст сокращался поочередно слева и справа. Если фрагмент содержал одно слово, то он не включался в реферат.

5. Результаты

Примеры работы системы

На рисунках приведены примеры работы системы – рефераты статьи «Системы автоматического реферирования» [4] без учета запроса (рис. 1) и с учетом двух разных запросов (рис. 2, 3), а также список десяти ключевых слов с наибольшим весом для каждого случая.

Хорошо видно, что простые шаблоны для обнаружения границ предложения дают сбой даже на таких «хороших» документах (предложения 4 и 5 на рис. 1; предложение 3 на рис. 2).

Системы автоматического реферирования

1. Поэтому методы создания и оценки рефератов должны развиваться параллельно.
2. Главное различие между средствами реферирования состоит в том, что они, по существу, формируют – краткое изложение или набор выдержек.
3. В средствах реферирования этого типа методы реферирования одного документа должны быть распространены на большой набор документов.
4. Средство реферирования мультимедиа, использует Broadcast News Navigator, который выполняет поиск, просмотр и реферирование теленовостей На экране представлен мультимедийный реферат информационного наполнения видеофрагмента, выданный на запрос поискового механизма.
5. 5 приведен примерный реферат, созданный системой Broadcast News Navigator [15] – средством поиска, просмотра и реферирования телевизионных новостей.

реферирование, система, автоматический, реферат, автоматическое реферирование, должный, текст, метод, средство, краткое изложение

Рис. 1. Реферат и список ключевых слов (в порядке убывания веса)

Системы автоматического реферирования

1. В качестве структур могут быть использованы формулы логики предикатов или такие представления, как семантическая сеть или набор фреймов.
2. В процессе преобразования концептуальное представление претерпевает несколько изменений.
3. 3 , этап синтеза одинаков для обоих подходов: текстовый генератор преобразует структурное или концептуальное представление в естественно-языковую аннотацию.
4. Средство реферирования полностью представляет семантическую информацию в виде связей между узлами в концептуальном графе, как таксономические (подкласс или экземпляр) или метонимические (часть) отношения.
5. Так как они опираются на формальное представление информационного наполнения документа, их можно настроить на весьма высокие степени сжатия, например, такие, которые требуются для рассылки сообщений на устройства PDA.

представление, семантический, семантическое представление, реферирование, система, автоматический, реферат, автоматическое реферирование, должный, текст

Рис. 2. Реферат по запросу «семантическое представление» и список ключевых слов (в порядке убывания веса)

Системы автоматического реферирования

1. Поэтому методы создания и оценки рефератов должны развиваться параллельно.
2. Первый (вверху) опирается на традиционный лингвистический метод синтаксического разбора предложений
3. И, наконец, разработчики средств реферирования все больше склоняются к гибридным системам, а исследователям все более успешно удается объединять статистические методы и методы, основанные на знаниях.
4. Целью методов оценки рефератов является определения адекватности (и достоверности) или пользы реферата по отношению к оригинальному тексту.
5. В средствах реферирования этого типа методы реферирования одного документа должны быть распространены на большой набор документов.

метод, лингвистический, лингвистический метод, реферирование, система, автоматический, реферат, автоматическое реферирование, должный, текст

Рис. 3. Реферат по запросу «лингвистические методы» и список ключевых слов (в порядке убывания веса)

Выделение словосочетаний

На рис. 4 представлен алфавитный список словосочетаний, выделенных из текста статьи «Системы автоматического реферирования» [4] по шаблону *A+N*. Как видно из рисунка, некоторые словосочетания являются шумовыми (перечеркнуты на рисунке), однако многие конструкции являются распространенными устойчивыми словосочетаниями (выделены жирным шрифтом).

***автоматическое реферирование**, аналитический этап, **весовой коэффициент**, **Геттисбергское обращение**, гибридный источник, **естественный язык**, идеальный реферат, избыточная информация, информационное наполнение, исходный текст, ~~каждый блок~~, ключевая фраза, ключевое предложение, концептуальная выжимка, **концептуальное представление**, **концептуальный подграф**, краткое изложение, критический реферат, онтологический справочник, основная мысль, поисковый механизм, **предметная область**, ~~различные источники~~, репрезентативная структура, семантическая информация, семантическое представление, **синтаксический разбор**, **статистическая важность**, **террористический акт**, характерный фрагмент, ~~широкий диапазон~~*

Рис. 4. Пример: список выделенных словосочетаний

Производительность

Разработанная нами система предназначалась в первую очередь для исследовательских целей, поэтому мы не ставили перед собой цели добиться высокой производительности.

Версия программы, которая использовалась для выполнения заданий дорожки контекстно-зависимого аннотирования РОМИП-2005, обработала около 21000 документов примерно за 60 часов на компьютере с процессором Pentium-4/2,4 ГГц (фактически использовались два компьютера сравнимой производительности, каждый работал примерно 30 часов). Время составления реферата сильно зависело от длины документа (~6 с для документа с исходным объемом 50 Кб, ~18 с – 100 Кб).

После этого мы оптимизировали программу, скорость работы возросла: ~1,2 с для документов объемом до 100 Кб, ~15 с для документов объемом 300 Кб; причем от 30% (для длинных документов) до 80% (короткие документы) времени занимает обработка текста с помощью *mystem*.

6. Оценка

РОМИП

Разработанная система *eXtragon* принимала участие в дорожке по контекстно-зависимому аннотированию РОМИП-2005.[3] Набор исходных данных, созданный на основе оценивавшихся запросов дорожек поиска по Веб-коллекции и по коллекции нормативно-правовых документов в рамках РОМИП-2004, включал около 21000 пар «документ+запрос». По каждой такой паре необходимо было сформировать аннотацию длиной не более 300 символов без HTML разметки. В ходе оценки асессоры оценивают релевантность документа запросу на основании аннотации. Результатом оценки является мера согласованности оценок для полных документов и для аннотаций.

Результаты оценки будут известны в августе 2005 года.

Сравнение выдач

Для оценки качества реферирования мы провели небольшое сравнительное исследование представления результатов поиска МП Яндекс и нашей системы. Мы взяли 10 случайных запросов из раздела

[3] См. http://www.romip.narod.ru/ru/2005/tracks/annotation.html

«Прямой эфир Яндекса» и для каждого запроса сформировали две страницы: (1) со сниппетами, полученными через службу Яндекс.XML[4], и (2) с рефератами, полученными с помощью системы *eXtragon*. Каждая страница содержала десять первых ссылок на документы (на один из запросов – с орфографической ошибкой – было возвращено только три ссылки, одна из которых оказалась неработающей), заголовки и до пяти фрагментов документа (в некоторых результатах, возвращенных Яндексом, фрагментов было меньше). Необходимо отметить, что фрагменты, получаемые через Яндекс.XML, отличаются от тех, которые видит пользователь Веб-интерфейса поисковой машины: в первом случае это целые предложения, во втором – контексты слов запроса. Страницы с рефератами Яндекса и *eXtragon* были оформлены одинаково, отличаясь только цветом фона.

Табл. 1. Результаты сравнения выдач

	Яндекс				eXtragon			
	1	2	3	**Σ**	1	2	3	**Σ**
1. Бегбедер	10	12	13	**35**	14	18	12	**44**
2. глина с опилками	6	20	16	**42**	15	20	12	**47**
3. Дальзавод	9	12	11	**32**	16	15	13	**44**
4. Иван Четвертый грозный	3	11	10	**24**	8	13	9	**30**
5. исстории про измены	2	5	4	**11**	2	0	0	**2**
6. как пройти собеседование	15	9	15	**39**	20	19	19	**58**
7. копии часов	20	16	20	**56**	20	15	19	**54**
8. металлургия Украины в 2005г	12	17	20	**49**	20	15	17	**52**
9. ниши внутри Торпедо	12	19	15	**46**	14	17	10	**41**
10. разработка веб сайтов	17	17	19	**53**	20	20	19	**59**
Итого	**106**	**138**	**143**	**387**	**149**	**152**	**130**	**431**

Каждому из трех участников были предъявлены 20 страниц (10 запросов × 2 системы). Мы попросили участников ответить на вопрос: «Можно ли по фрагментам составить представление о документе и оценить его соответствие запросу?» Шкала оценки выглядела следующим образом: 0 – непонятно; 1 – что-то есть; 2 – все понятно. Результаты представлены в табл. 1. Два из трех участников

[4] См. http://xml.yandex.ru

оценили выше рефераты *eXtragon*, при этом все трое отметили, что часто эти рефераты были слишком большими и включали в себя элементы оформления страницы (элементы навигации, меню, указания на авторские права и т.д.).

Сравнение на основе исследования поведения пользователей

Дополнительно мы предприняли попытку провести эксперимент по сравнению двух систем: (1) Яндекс.XML и (2) Яндекс.XML+ *eXtragon*, аналогичный описанному в [11]. Для этого мы разработали унифицированный интерфейс (рис. 5). Интерфейс одинаково (различается только цвет фона) отображает результаты поиска, полученные от (1) Яндекс.XML, и (2) те же результаты с рефератами, сформированными с помощью *eXtragon*.

Схема эксперимента аналогична описанной в [6]. В эксперименте приняли участие четыре человека. Каждый участник выполнял четыре задания по фактографическому поиску (рис. 5) в соответствии с табл. 2. Перед выполнением заданий участники заполняли анкеты (имя, возраст, пол, образование, профессия, навыки Веб-поиска). После выполнения участники оценивали полезность, информативность и длину рефератов каждой из систем. Дополнительно регистрировались действия пользователей (запросы, количество просмотренных страниц с результатами поиска, обращения к оригинальным документам).

Табл. 2. Соответствие «задание – участник – система» (Y – Яндекс, X - eXtragon)

Участник	Задания			
	1	2	3	4
1	Y	Y	X	X
2	X	Y	Y	X
3	X	X	Y	Y
4	Y	X	X	Y

1. Когда и кем изобретен трехфазный асинхронный двигатель?
2. Координаты города Урюпинска?
3. Где и когда родился министр финансов СССР 1989-1991 гг.?
4. Как звали последнего мужа сестры жены французского поэта Арагона?

Рис. 5. Задания для фактографического поиска

Если судить по оценкам участников и протоколам их действий, сравниваемые системы различаются незначительно. Как и в предыдущем тесте, отмечалась «зашумленность» рефератов *eXtragon* эле-

ментами оформления страниц. К сожалению, технические и организационные ошибки, допущенные в ходе эксперимента, а также небольшое количество участников не позволяют в настоящий момент делать уверенных выводов на основе полученных данных.

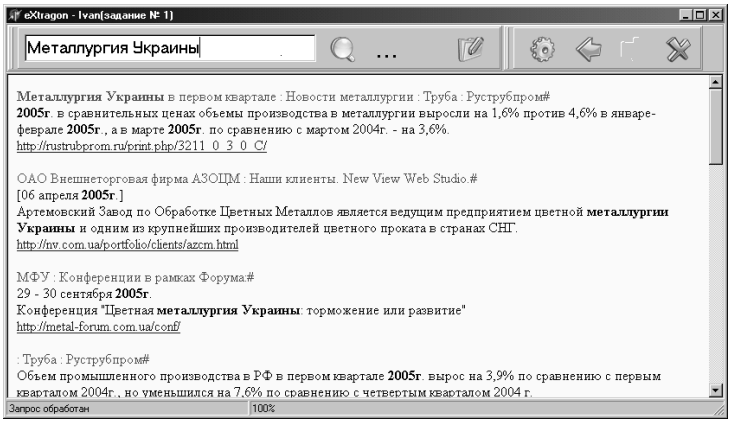

Рис. 6. Унифицированный интерфейс для сравнительного анализа систем

7. Выводы и обсуждение результатов

Предварительные результаты позволяют говорить о приемлемом качестве реферирования Веб-документов. Мы сможем сделать более обоснованные заключения об эффективности использования предложенного подхода для задач информационного поиска после получения результатов РОМИП.

Основное преимущество описанного метода извлечения предложений для задачи интернет-поиска – это возможность составлять рефераты текстовых документов, даже если в них отсутствуют слова запроса. Такая ситуация может возникать, если документ найден по тексту ссылки или слово запроса присутствует только в заголовке, подписи к картинке, в разделе метаописаний «ключевые слова». Важно и то, что метод универсален и не накладывает ограничений на тематику и стиль документов.

Эксперименты с разработанной системой показали, что основным недостатком является смешивания основного содержания реферируемых страниц и элементов оформления при обработке. Кроме того, простые шаблоны, которые использовались для определения границ предложений, дают сбой на многих Веб-документах (это справедливо и для «хороших» документов, см. рис. 1, 2). Дополни-

тельно, при реферировании больших документов представляется перспективным использование информации об их внутренней структуре.

В использованной нами схеме вычисления весов предложений выделенные словосочетания относительно слабо влияли на результат. Нам представляется, что задача учета устойчивых словосочетаний в процессе реферирования может представлять специальный интерес.

Предложенный метод выделения важных предложений текста может быть эффективно реализован в современных информационно-поисковых системах в том случае, если их индексные структуры хранят информацию о позициях слов с привязкой к предложениям.

Благодарности

Мы благодарим компанию «Яндекс» за поддержку проекта. Мы признательны всем, кто оказал содействие при тестировании системы и сравнительной оценке качества реферирования.

Литература

1. Губин М.В., Меркулов А.И. Эффективный алгоритм формирования контекстно-зависимых аннотаций // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог'2005» (Звенигород, 1-6 июня 2005 г.). – М.: Наука, 2005. – С. 116–120.
2. Солтон Дж. Динамические библиотечно-информационные системы. – М.: Мир, 1979.
3. Ступин В. С. Система автоматического реферирования методом симметричного реферирования // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог'2004». («Верхневолжский», 2-7 июня 2004 г.). – М.: Наука, 2004. – С. 579-591.
4. Хан У., Мани И. Системы автоматического реферирования // Открытые системы, 2000. – №12. Эл. версия: http://www.osp.ru/os/2000/12/067_print.htm
5. Brandow R., Mitze, K., & Rau, L. F. (1995). Automatic condensation of electronic publications by sentence selection. In *Information Processing & Management*, 31 (5), 675-685.
6. Braslavski P., Shishkin A. A User-Centered Comparison of Web Search Engines. In *Computational Linguistics and Intelligent Technologies. Proceedings of the Dialogue'2005 conference*. Zvenigorod, June 1-6, 2005. P. 554-560.

7. Iatsko V. (2001). Linguistic Aspects of Summarization. In *Philologie im Netz*, 18, 33-46. Available online: http://www.fu-berlin.de/phin/phin18/p18t3.htm
8. Manning C. D., Schütze H. Foundations of Statistical Natural Language Processing. MIT Press, 2000.
9. Nomoto, T., & Matsumoto, Y. (2003). The diversity-based approach to open-domain text summarization. In *Information Processing & Management*, 39, 363-389.
10. Tombros, A. & Sanderson, M. (1998). Advantages of Query Biased Summaries in Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, August 24-28, 1998, Melbourne, Australia, 2-10.
11. White, R. W., Jose, J. M., & Ruthven, I. (2003). A task-oriented study on the influencing effects of query-biased summarisation in web searching. In *Information Processing & Management*, 39, 707-733.
12. Yang, Ch. C. & Wang, F. L. (2003). Fractal Summarization for Mobile Devices to Access Large Documents on the Web. In *Proceedings of the WWW2003*, May 20-24, 2003, Budapest, Hungary. Available online: http://www2003.org/cdrom/papers/refereed/p681/p681-yang-html/p681-yang.html

Automatic Query-Biased Summarization of Web documents

Pavel Braslavski, Ivan Kolychev

This report describes basic principles and implementation of an experimental summarization system for Web documents. The system employs a *sentence extraction* approach dated back to the 50-60s, with minor modifications. Sample summaries and system performance estimates are presented. Different frameworks for summarization evaluation are described. Preliminary results prove acceptable quality of the obtained summaries. The proposed approach can be effectively implemented in a modern information retrieval system if the search index stores word positions related to the sentence level.