Вычисление значимой части текста (в поисково-аналитической системе «Галактика-ZOOM»)

Антонов А.В., Курзинер Е.С.
Корпорация «Галактика»

В докладе исследуется алгоритм вычисления значимого отрывка документа, который применяется в поисково-аналитической системы «Галактика-ZOOM», работающей с лексически необработанными текстовыми базами данных.
Основная аналитика системы заключается в определении наиболее значимых слов выборки, то есть характеризующих контекст запроса в отличие от контекста всей базы. На основе такого «информационного портрета» возможно осуществление многих дополнительных функций: уточнение запроса; ранжирование документов выборки (по их соответствию весомой части инфопортрета); ранжирование выборки по «чужому» инфопортрету. Последняя функция фактически создает исходному запросу некоторое семантическое или стилистическое поле, определяемое этой «чужой» выборкой - и увеличивает информационный вес документов, соответствующих эталонному, «чужому», набору значимых слов.
Одна из функций системы – выделение в документе наиболее значимого отрывка (посредством последовательного ранжирования равных (то есть содержащих одинаковое количество слов) частей документа по информационному портрету выборки, с учетом коэффициента значимости «заголовочной» части текста). Представляемая функция позволяет пользователю просматривать выборку, не раскрывая конкретные документы.

Одной из актуальных задач поисковых систем является оптимизация отображения найденной информации, то есть выбора наиболее соответствующей *запросу* части документа. Однако не менее важным оказывается и характерность показываемого отрывка *теме документа*. Большинство же существующих систем показывают в найденных по запросу документах отрывок текста, содержащий слова запроса, хотя понятно, что не всегда лучший ответ повторяет вопрос. Одним из решений подобной проблемы является использование метода кластеризации слов (см. [12])[1] . Мы же больше склоняемся к также используемому в практике методу вычисления веса значимых слов. Предметом нашего исследования было выяснение характеристик текста документа и запроса, влияющих на подобную оптимизацию вычисления наиболее «адекватного» отрывка. Основой исследования стала гипотеза о возможности корреляции веса значимого отрывка в соответствии с весом начальной части документа (см. подобные идеи – о важности заголовка - в [6], [13]). Исследование проводилось в рамках работы поисково-аналитической системы «Галактика-ZOOM».

[1] Обзор работ по теме “automatic summarization” – см. [5], [6].

Поисково-аналитическая система «Галактика-Зум» имеет дело с большими текстовыми базами данных, не прошедшими никакой предварительной лексической обработки (см. [1]-[3]) . На конкретный запрос происходит отбор соответствующих документов с параллельным построением так называемого «информационного портрета», то есть выделением значимых для данной выборки слов или словосочетаний. При этом следует подчеркнуть, что лексически (и информационно) инфопортрет выборки представляет собой, по сути, контекст запрашиваемого текста. На основе информационного портрета выборки системой реализуется ряд полезных функций: автоматическое ранжирование документов выборки, полуавтоматическое уточнение запроса (маркированием интересующих пользователя слов и словосочетаний в полученном инфопортрете), автоматическое уточнение запроса путем ранжирования по «чужому» инфопортрету (что фактически усиливает определенное семантическое или стилистическое поле выборки)

Представленная в этом докладе возможность оптимизации отображения «запросного» отрывка текста также основывается на информационном портрете выборки: наиболее актуальный для запроса отрывок конкретного документа определяется наибольшим суммарным весом значимых (то есть «инфопортретных») слов. Вес же вычисляется последовательно для всех потенциальных «равнодлинных» отрывков текста, с учетом так называемого «заголовочного» коэффициента, являющегося некой постоянной величиной, лишь корректирующей альтернативный выбор «лучшего» отрывка.

Попытаемся проанализировать вышеизложенное на примерах нескольких простых запросов, различающихся прежде всего своими контекстами, а формально – информационными источниками (изданиями), что оказывает существенное влияние на выбор значимого отрывка. Похоже, что на «качественность» документа влияют цельность[2] запроса («Макаревич и Гребенщиков» vs. «Макаревич»), вхождение в длинный парадигматический ряд (как правило, это имена собственные – «Макаревич»). Соотношение суммарной значимости «заголовочной» (содержащей заголовок) и «значимой» (содержащей значимые слова с наибольшим суммарным весом) частей документа отражает однородность или разнородность, а также существенность или несущественность документа для данного запроса.

Исследование проводилось по базе печатных СМИ методом экспертной оценки начального отрывка документа выборки по следующей шкале: «удовлетворяет запросу и теме документа», «не удовлетворяет запросу или теме документа», «совпадает с наиболее весомым отрывком». В каждой из выборок (было проведено 12 запросов) рассматривалось по 20 документов.

Пример 1.

Запрос: Андрей Макаревич

(4026 документов в полученной выборке)

Запрос цельный, и поэтому заголовочная часть документа почти всегда оказывается более информативной, удовлетворяя одновременно теме запроса и теме документа.

Инфопортрет полученной выборки[3]:

Коэффициент значимости

[2] Цельный запрос - представляющий собой лексическое (денотативное) целое: Андрей Макаревич - это одна единица, гороскоп и питание – две.

[3] Точнее, его верхушка.

```
  67.5    МАКАРЕВИЧ
  19.9    АНДРЕЙ
  10.9    ПЕСНЯ
  9.68    КОНЦЕРТ
  8.64    МУЗЫКАНТ
  8.3     НОВЫЙ АЛЬБОМ
  8.18    СМАК
  8.01    АЛЬБОМ
  6.57    КРЕОЛЬСКОЕ ТАНГО
  4.66    МУЗЫКА
  4.64    РУССКИЙ РОК
  4.4     ЯРМОЛЬНИК
  4.37    ХРУСТАЛЬНАЯ ТУРАНДОТ
  3.89    ТИХИЙ ОМУТ
```

Однако имя собственное обычно ассоциируется с длинным рядом других имен собственных, что обычно снижает информативность отобранных документов, а соответственно, и отбираемого отрывка в виду неоднородности контекста. Отсюда же может вытекать и косвенность отношения начального отрывка к запросу.
Наиболее значимый отрывок первого документа выборки, полученной по этому запросу:

1. Итоги (Москва) , N012 (21.3.2000) 

... **АЛЬБОМ МАКАРЕВИЧА**

1986 - ГРУППА ГОТОВИТ ДВОЙНУЮ **ПЛАСТИНКУ** "РЕКИ **И** МОСТЫ"

1987 - В ПРОДАЖУ ПОСТУПАЕТ ПЕРВЫЙ ДИСК-ГИГАНТ "В ДОБРЫЙ ЧАС"

1989 - В "**ЛУЖНИКАХ**" ПРОХОДИТ **ЮБИЛЕЙНЫЙ КОНЦЕРТ**, ПОСВЯЩЕННЫЙ ДВАДЦАТИЛЕТИЮ **АНСАМБЛЯ**

1990 - В МОСКВЕ ОТКРЫВАЕТСЯ **ВЫСТАВКА** ГРАФИКИ **•АНДРЕЯ •МАКАРЕВИЧА**

1992 - **МАКАРЕВИЧ ВЫПУСКАЕТ КНИГУ** "**ВСЕ** ОЧЕНЬ **ПРОСТО**"

1993 - В ГКЦЗ "РОССИЯ" **ПРАЗДНУЕТСЯ** СОРОКАЛЕТИЕ **МАКАРЕВИЧА**

1994 - "МАШИНЕ" 25 **ЛЕТ**. **МАКАРЕВИЧ** ДАЕТ РЯД **СОЛЬНЫХ КОНЦЕРТОВ** В МОСКВЕ

1997 - ПРОВОДИТ ТРИ СОВМЕСТНЫХ С **ГРЕБЕНЩИКОВЫМ КОНЦЕРТА** В "РОССИИ" ПОД ДЕВИЗОМ "**ДВАДЦАТЬ ЛЕТ** СПУСТЯ"

1998 - "МАШИНА" **ВЫПУСКАЕТ ПЛАСТИНКУ "ОТРЫВАЯСЬ"**

1999 - ГРУППА ОТМЕЧАЕТ **ТРИДЦАТИЛЕТИЕ ГАСТРОЛЬНЫМ ТУРОМ И ВЫПУСКАЕТ** НОВЫЙ **АЛЬБОМ** (29491/61061) [4] ...

Начальный отрывок:
SOURCE: Итоги (Москва) , N012
DATE: 21.3.2000
MESSAGE: ЖИВОТ **АРХИТЕКТОРА**.

1953-11 ДЕКАБРЯ **РОДИЛСЯ •АНДРЕЙ •МАКАРЕВИЧ**

1969 - 27 ИЮНЯ ОРГАНИЗОВАНА ГРУППА "МАШИНА ВРЕМЕНИ"

1971 - СОСТОЯЛСЯ ПЕРВЫЙ **ПУБЛИЧНЫЙ КОНЦЕРТ** "МАШИНЫ" В **ДК** "ЭНЕРГЕТИК"

[4] Первое число – суммарный вес слов начального отрывка документа, второе число – суммарный вес слов наиболее значимого отрывка.

1976 - "МАШИНА ВРЕМЕНИ" УЧАСТВУЕТ В ТАЛЛИНСКОМ **ФЕСТИВАЛЕ ПОПУЛЯРНОЙ МУЗЫКИ**

1977 - ГРУППА **СНИМАЕТСЯ** В **ДОКУМЕНТАЛЬНОМ ФИЛЬМЕ** "ШЕСТЬ ПИСЕМ О БИТЕ"

1978 - **ВЫХОДИТ АЛЬБОМ** "ДЕНЬ **РОЖДЕНИЯ**"

1980 - **ПРЕМЬЕРА** НА **ТЕЛЕВИДЕНИИ** В "ГОЛУБОМ **ОГОНЬКЕ**", ГРУППА УЧАСТВУЕТ В **ФЕСТИВАЛЕ** "ТБИЛИСИ-80" (ПЕРВАЯ **ПРЕМИЯ** ПЛЮС СПЕЦИАЛЬНЫЙ **ПРИЗ** СОЮЗА ЖУРНАЛИСТОВ ЗА **ТЕКСТЫ**)

1983 - **ЗАПИСЫВАЕТСЯ СОЛЬНЫЙ АКУСТИЧЕСКИЙ АЛЬБОМ МАКАРЕВИЧА**

1986 - ГРУППА ГОТОВИТ ДВОЙНУЮ **ПЛАСТИНКУ** "РЕКИ **И** МОСТЫ"

1987 - В ПРОДАЖУ ПОСТУПАЕТ ПЕРВЫЙ ДИСК-ГИГАНТ "В ДОБРЫЙ ЧАС"

1989 - В "**ЛУЖНИКАХ**" ПРОХОДИТ **ЮБИЛЕЙНЫЙ КОНЦЕРТ**, ПОСВЯЩЕННЫЙ ДВАДЦАТИЛЕТИЮ **АНСАМБЛЯ**

1990 - В МОСКВЕ ОТКРЫВАЕТСЯ **ВЫСТАВКА** ГРАФИКИ **•АНДРЕЯ •МАКАРЕВИЧА**

1992 - **МАКАРЕВИЧ ВЫПУСКАЕТ КНИГУ** "**ВСЕ** ОЧЕНЬ **ПРОСТО**"

1993 - В ГКЦЗ "РОССИЯ" **ПРАЗДНУЕТСЯ** СОРОКАЛЕТИЕ **МАКАРЕВИЧА**

1994 - "МАШИНЕ" 25 **ЛЕТ**. **МАКАРЕВИЧ** ДАЕТ РЯД **СОЛЬНЫХ КОНЦЕРТОВ** В МОСКВЕ

1997 - ПРОВОДИТ ТРИ СОВМЕСТНЫХ С **ГРЕБЕНЩИКОВЫМ КОНЦЕРТА** В "РОССИИ" ПОД ДЕВИЗОМ "**ДВАДЦАТЬ ЛЕТ** СПУСТЯ"

1998 - "МАШИНА" **ВЫПУСКАЕТ ПЛАСТИНКУ** "**ОТРЫВАЯСЬ**"

1999 - ГРУППА ОТМЕЧАЕТ **ТРИДЦАТИЛЕТИЕ ГАСТРОЛЬНЫМ ТУРОМ И ВЫПУСКАЕТ** НОВЫЙ **АЛЬБОМ**

2000 - ВМЕСТЕ С ГРУППОЙ "**ВОСКРЕСЕНИЕ**" "МАШИНА ВРЕМЕНИ" ПУСКАЕТСЯ В ОЧЕРЕДНОЙ **ЮБИЛЕЙНЫЙ ТУР** ПОД НАЗВАНИЕМ "50 **ЛЕТ** НА ДВОИХ"

Пример 2.

Запрос: Гороскоп и питание

(795 документов в выборке)

Формально являясь нецельным запросом, но, обладая общим для конкретных запросных слов контекстом, этот запрос, как и в первом примере, по большей части, хорошо удовлетворяется начальными отрывками документов.

Кажется, что принадлежность запросного слова к длинному парадигматическому ряду является как существенным недостатком соответствующего документа, так и препятствием выделения значимого отрывка в документе. Поэтому, вероятно, «заголовочный» коэффициент должен быть «коэффициентом цельности».

Инфопортрет:

Коэффициент значимости	
2.62e+003	ГОРОСКОП
2.17e+003	ТЕЛО
1.88e+003	ОБР
1.77e+003	КВ
1.61e+003	СОСТ

```
 1.17e+003 Р-Н
 1.07e+003 ВОДОЛЕЙ
 1.05e+003 КОЗЕРОГ
 1.01e+003 СКОРПИОН
 1.01e+003 ОВЕН
   940     СТРЕЛЕЦ
   931     ПИТАНИЕ
   867     ДЕВА
   855     БЛИЗНЕЦ
   757     ВЕСЫ
   735     СОТКА
   720     НЕДОРОГО
   684     КИРП
   668     ОТР
   652     РЫБА
   598     ПОСТАРАТЬСЯ
   530     ТЕЛЕЦ
   499     АСТРОЛОГ
   489     ПОЗНАКОМИТЬСЯ
   472     РУБ
   452     БЛАГОПРИЯТНЫЙ
   397     Ч
   385     ЯО
   378     М
   363     АСТРОЛОГИЯ
```

Наиболее значимый отрывок первого наиболее значимого документа выборки характеризует контекст запроса:

Провинциальный репортер (Липецк) , N045 (5.11.2003)

... 350 **руб.** Освоскоп - 150 **руб.** **Терморегулятор** (точность 0,2 С) - 320 **руб.** Инструкция. **Гарантия.** Выписывайте: **Обр.:** 460006, г. Оренбург, **ул.** Гусева-32, **"Инкубатор".**

Одноразовая **посуда** от производителя. Ищем **партнеров** по сбыту. ООО "Пласт-С". **Обр.:** г. **Саратов.** * (8452) 59-86-42; 73-41-35; **факс** 43-36-18.

СПК "Рыбколхоз" им. И.В. Абрамова" реализует прудовую **рыбу:** живую, охлажденную, мороженую **оптом** и в **розницу.** Цена **договорная. Обр.:** Ростовская **обл.,** г. Семикаракорск, **пр.** Абрамова, 1. т.(86356) 2-25-01, 2-14-30. **Факс** 2-19-65.

Инкубатор на 70 **яиц** - от **автоаккумулятора** 12В и **сети.** Наложенным **пл.** 970 **руб.,** работающий от **сети** - 750 **руб. Терморегулятор** - 300 **руб. Розница** и **опт. Обр.:** 347900, г. **Таганрог** (7554/48194) ...

Огромная разница суммарной значимости начального и «значимого» отрывка свидетельствует о неоднородности документа. А значимым для запроса он оказался в виду значимости контекста.

Так же неинформативно и начало документа:

Начало документа (сводка объявлений):

SOURCE: Провинциальный репортер (Липецк) , N045

DATE: 5.11.2003

MESSAGE: Услуги.

Ремонт холодильников Ремонт холодильников на дому, **установка** моторов. Выдается **гарантия**. т.41-28-56. 40-80-80 **Св-во** 03-3, N20640

Ремонт холодильников всех марок, в том числе "**Стинол**", замена моторов,

упл. **резины**. **Выезд** в район, **гарантия**. т.47-19-47, 38-72-24. **Св**. 2361 от 23.05.95 **Адм**. г.**Липецка**

Ремонт холодильников всех марок и "**Стинол**" на дому без **выходных**, по городу и области. **Гарантия** до 5 **лет**. т.47-92-74. **Св-во** 03-2 NN119-83 от 27.09.98 г. **ИНН** 482400029962.

Ремонт видеоаппаратуры и **сотовых** телефонов

Телемастер. Качественный **ремонт телевизоров**, видеомагнитофонов, музыкальных центров. **Умеренные** цены. **Гарантия** до года. т.45-05-36. **Св**.0682 NN11013

Ремонт любых **телевизоров**, вызов **бесплатный** в любом районе. **гарантия** 2 года. т.43-56-75, 8-910-353-73-52. **Св**. 03-3 N6963 от 25.12.95 **Адм**. г.**Липецка**

Следующий документ:

Молва (Владимир) , N036 (19.3.2002)

... следующих советов. **ОВЕН**. Первая **декада**: 21 - 31.03. Вторая **декада**:1 - 10.04. Третья **декада**: 11 - 20.04. **Овны** не очень разборчивы в **еде** и, как правило, **равнодушны** к тому, что **едят**. В **зрелые** годы **Овну желательно** сократить количество принимаемой **пищи** в связи с уменьшением **активности**. Следует **избегать жирной пищи**, вызывающей атеросклероз. Любимые **продукты Овна** - ягнятина, баранина или козлятина. **Общительный Овен** любит **хорошо** провести время с друзьями и, очевидно, является большим ценителем пива. Это не случайно, поскольку одним из **растений Овна** является хмель. Кстати, пиво не **принесет** вреда **родившемуся** под **знаком Овна**. **Овощи Овна - морковь** (31977/37998) ...

Соотношение суммарной значимости отрывков свидетельствует об однородности документа и его соответствию теме запроса.

Начальный отрывок:

SOURCE: Молва (Владимир) , N036
DATE: 19.3.2002
MESSAGE: КАК **ПРАВИЛЬНО ПИТАТЬСЯ** ПО **•ГОРОСКОПУ**.

Многие ли из нас знают, что каждому из **знака Зодиака** соответствуют определенные виды **пищи**? Если у вас **часто** бывает плохое **настроение** или головные **боли**, вы **чувствуете** себя неуютно, то это происходит из-за того, что вы **употребляете продукты**, не характерные **для** вас. **Астрологи** предлагают вам это **исправить** с помощью следующих советов. **ОВЕН**. Первая **декада**: 21 - 31.03. Вторая **декада**:1 - 10.04. Третья **декада**: 11 - 20.04. **Овны** не очень разборчивы в **еде** и, как правило, **равнодушны** к тому, что **едят**. В **зрелые** годы **Овну желательно** сократить количество принимаемой **пищи** в связи с уменьшением **активности**. Следует **избегать жирной пищи**, вызывающей атеросклероз. Любимые **продукты Овна** - ягнятина, баранина или козлятина. **Общительный Овен** любит **хорошо** провести время с друзьями и, очевидно, является большим ценителем пива. Это не случайно, поскольку одним из **растений Овна** является хмель. Кстати, пиво не **принесет** вреда **родившемуся** под **знаком Овна**. **Овощи Овна** - **морковь**, хмель, репчатый лук, **перец сладкий** и стручковый горький, редис, лук-шалот. **Фрукты** - грейпфрут и арбуз. **Овны** любят **все острое**, **особенно** приправы с луком и **перцем**. **Травы** и **специи Овна** первой **декады** - чилийский **перец**, кэрри, **чеснок**, **хрен**, **горчица**, лук, базилик, кориандр, тмин, имбирь, перечная **мята**, анис, гвоздика, клен, мускатный **орех**, **шалфей**. **Травы** и **специи Овна** второй **декады** имеют более **тонкий** и изысканный аромат - розмарин, шафран, кунжут, цикорий, цитрон, женьшень, лавровый лист.**Овнам** третьей **декады подойдут**

Результат исследований подтвердил нашу гипотезу о необходимости введения так называемого «заголовочного» коэффициента: примерно в 70% начальный отрывок документа удовлетворяет как запрос, так и тему документа, в 15% случаев начало оказывается лучше, и только в 15% - отражая тему документа, начало не отвечает запросу. При этом заголовочный отрывок оказывается информативней, если документ однороден и посвящен теме запроса, и хуже, если документ представляет собой некие сводки или списки, то есть не является однородным. Исследование соотношения суммарной значимости слов значимого (в соответствии с инфопортретом) отрывка – и начального, заголовочного, отрывка, привел также к выводу о возможности выявления неоднородности, или нецельности, «неповествовательности», текста документа - при большом разрыве в суммарной значимости этих отрывков.

Литература

1. Антонов А. Большие Информационные Объекты//Сб. ВИНИТИ №4, 2001.
2. Антонов А. Информационно-поисковая система «Galaktika-ZOOM» с элементами анализа на гипермассивах информации//Сб. ВИНИТИ №8, 2001.
3. Антонов А., Мешков В. Современные проблемы поисковых систем и некоторые пути их преодоления//Сер. «Аналитика-Капитал», Москва, 2000.
4. Thomas Hofmann. Probabilistic latent semantic indexing. In Proc of SIGIR'99, pp.50-57, 1999.
5. Weizheng Gao. A survey of text Summarization. http://flame.cs.dal.ca/~wgao/6906project.pdf
6. Text summarization and Information Extraction A survey of algorithms and tools. http://www-2.cs.cmu.edu/~madhavi/11-742/report.pdf
7. T.R.Lynam, C.L.A.Clarke, G.V.Cormack. Information Extraction with Term Frequencies//Computer Science University of Waterloo, Ontario, Canada. http://acl.ldc.upenn.edu/H/H01/H01-1036.pdf.
8. Klaus Zechner. Automatic Text Abstracting by Selecting Relevant Passages.MSc Dissrtation in Cognitve Science and Natural Language. August 12, 1995. http://www-2.cs.cmu.edu/~zechner/abstr.ps.
9. Fernando Llopis, Jose Luis Vicedo, Antonio Ferrandez. Passage Selection to Improve Question Answering. http://www.isi.edu/~cyl/wsqa-coling2002/papers/P0011.pdf.
10. Tatsunori Mori and Takuro Sasaki. Information Gain Ratio meets Maximal Marginal Relevance. A method of Summarization for Multiple Documents//Proceedings of the Third NTCIR Workshop. Sep.2001-Oct.2002. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-TSC-MoriT.pdf.
11. Martin Hassel. A summary of the Paper A Robust Practical Summarizer//GSLT Information Access Course. http://www.nada.kth.se/~xmartin/kurser/gslt-ia/textsumsum.pdf
12. Dragomir R.Radev, Hongyan Jing, Malgorzata Budzikowska. Centroid-based summarization of multiple documents sentence extraction, utility-based evaluation, and user studies. http://arxiv.org/ftp/cs/papers/0005/0005020.pdf
13. Mark T. Maybury, Inderjeet Mani. Automatic Summarization. Tutorial Notes. American/European Conference on Computational Linguistics. Toulouse. France. 8 July 2001
http://www.mitre.org/tech/itc/maybury/summarization/manimayburysummarization.pdf