ОПРЕДЕЛЕНИЕ ТЕМАТИЧЕСКИ ЗНАЧИМЫХ ДОКУМЕНТОВ В СИСТЕМЕ ГАЛАКТИКА-ZOOM (АВТОРУБРИКАЦИЯ)

А.В. Антонов , Е.С. Курзинер

Корпорация «Галактика», Москва

alexa@galaktika.ru koorz@galaktika.ru

В статье описывается методика авторубрикации, используемая в поисково-аналитической системе «Галактика-Зум». Предварительно системой определяются информационные портреты, или ключевые темы, конкретных рубрик - по оригинальной технологии выделения и ранжирования ключевых тем. Затем автоматически происходит автоклассификация документов методом сравнения информационных портретов документа и заданных рубрик.

Задача автоматического разнесения информационного потока по тематическим рубрикам является, бесспорно, одной из важнейшей в области обработки информации, актуальной, но не новой: предпринимались различные попытки авторубрикации документов по ключевым словам, заголовкам и т.п. Мы также пытаемся решать эту проблему, исходя из возможностей и преимуществ нашей системы.

Поисково-аналитическая система «Галактика-Зум», осуществляя поиск информации в текстовой базе, определяет так называемый «информационный портрет» запрашиваемой темы, то есть набор упорядоченных по значимости ключевых слов и словосочетаний, характерный именно для данной выборки. Подобная упорядоченность отражает ранг частотности ключевой темы (слова или словосочетания) выборки на фоне этой же темы в целой базе. Такая технология позволяет решать, в частности, задачу ранжирования документов выборки по значимости – фактически по наибольшему соответствию инфопортрету выборки количества значимых тем и их ранга в рассматриваемом документе. Эту же технологию мы используем и для решения задачи автоклассификации, или авторубрикации, документов – на основе сравнения инфопортрета классифицируемого документа с эталонным

веса (значимости) эталонных слов. Таким образом фактически происходит ранжирование документов по соответствию эталонному инфопортрету. Определенная «пороговая» значимость, величина которой была определена экспериментальным путем, отсекает незначительно похожие, документы, незначимые для рубрики, то есть недостоверно относящиеся к ней. В принципе, таким образом устраняется и проблема омонимии. Так что опция коррекции эталонного инфопортрета больше нужна для сужения тематики и стиля отбираемых документов.

Решение задачи автоматической рубрикации, или автоклассификации, документов в системе инфопортретом рубрики, с отсечением документов, соответствие инфопортретов которых эталонному ниже определенного порога.

Эталонный инфопортрет конкретной рубрики может представлять собой либо реальный инфопортрет выборки, полученной по адекватно сформулированному запросу, либо этот же инфопортрет, откорректированный «вручную». Под такой коррекцией понимается *изменение* пользователем *веса* ключевых слов и/или *удаление* их из инфопортрета выборки, а также *добавление* своих. Необходимость подобного вычищения связана с тем, что даже на самый удачный запрос в инфопортрет могут попасть и омонимы, и слова, имеющие омонимичные с «чужеродными» словами формы. И то и другое будет ошибочно отсылать документы в «омонимичную» рубрику. Кроме того, функция редактирования инфопортрета позволяет сужать тематику и стиль отбираемых документов, что особенно важно при слишком большом объеме информации.

Сравнение инфопортрета конкретного документа с эталонным инфопортретом происходит на основе нескольких критериев: количества «эталонных» (то есть попавших в эталонный инфопортрет) слов; соотношения количества и веса «эталонных» слов с общим количеством слов в инфопортрете документа;

«Галактика-Зум» сводится, таким образом, к следующему:

1. Создание рубрик, определение инфопортрета каждой рубрики.
2. Коррекция полученных инфопортретов.
3. Сравнение инфопортрета документа с инфопортретами рубрик, с отсечением малохарактерных для рубрики документов.

Посмотрим, как функция автоклассификации работает в «Галактике-Зум» на базе СМИ. Одна из основных задач анализа подобной базы – определение значимых тем по дате и разнесение этих значимых тем «по интересам». Мы решили

проанализировать выборку, полученную на запрос date: 01.09.2004.

Так выглядела наиболее значимая часть, верхушка инфопортрета этого дня:

УЧЕБНЫЙ ГОД
ПРОПОРЦИОНАЛЬНАЯ СИСТЕМА
ФРАНЦУЗСКИЙ ЖУРНАЛИСТ
МУЖАХОЕВА
ТЕРАКТ
МОСКОВСКАЯ ОБЛАСТЬ
ШИРАК
ИСЛАМСКАЯ АРМИЯ
УСТАЛОВ
НАГАЕВА
ВЕНЕЦИАНСКИЙ ФЕСТИВАЛЬ
ПАРТИЙНЫЙ СПИСОК
ЦИК
ТУ-134
ШРЕДЕР
ФЕДЕРАЛЬНОЕ АГЕНТСТВО
РЕСПУБЛИКАНСКАЯ ПАРТИЯ
УЧЕБНЫЙ
ВЕНЕЦИЯ
ВЗРЫВ
ЗАСТРОЖНАЯ
ЕВЛАПОВА
ШАХИДКА
АЛХАНОВ
ФРАНЦУЗСКИЙ ЗАЛОЖНИК
ДЖЕБИРХАНОВА
ВЕНЕЦИАНСКИЙ
ФСТ
УДАЛЕННЫЙ ДОСТУП
СМЕРТНИЦА

Как видно, инфопортрет получился довольно разношерстным, хотя главными темами дня были теракты в самолетах накануне и новый учебный год. Однако мы искали.новости «по интересам». Для этого была проведена автоматическая классификация документов выборки по следующим рубрикам:

1. алкоголизм-наркомания
2. беременность
3. диета (без рекламы)
4. ДТП
5. жилье
6. комнатные растения
7. кошки
8. пожар
9. религия

В каждую рубрику отправлялось не больше 25 документов, при условии, что степень соответствия документа информационному портрету рубрики составляет не менее 60%. Такая величина «порога достоверности» была выбрана для большей точности, исходя из результатов эксперимента, проведенном в ходе участия в семинаре РОМИП [2]. Достоверность попадания нужных документов в рубрику в среднем составляет 80-90%, в ущерб, конечно, полноте информации, однако мы исходили из того, что для большого массива данных такая мера является оптимальной.

В результате нашего эксперимента из 1029 документов выборки в представленных рубриках оказалось следующее количество документов:
1) 25; 2) 11; 3) 25; 4) 25; 5) 3; 6) 25; 7) 8; 8) 16; 9) 25.

Далее экспертами оценивалась релевантность попадания конкретного документа в ту или иную рубрику. Из 163 документов, распределившихся по указанным девяти рубрикам, релевантными были признаны 134 документа, то есть 83%. Это при очень строгой оценке: не считались релевантными документы, имеющие косвенное отношение к проблеме рубрики – в которых ключевые темы выборки не были ключевыми для данного документа. При нестрогой же экспертной оценке, то есть при включении в инфопортрет документа смежных тем, релевантность достигла 100%.

Такой результат, как нам кажется, подтверждает правильность технологии автоматической классификации документа, используемой в поисково-аналитической системе «Галактика-Зум» – технологии рубрикации документов по информационным портретам, отражающим ключевые темы конкретных рубрик. При наличии в инфопортрете однозначно относящихся к данной рубрике ключевых тем наблюдается стопроцентное попадание в нее релевантных документов; при наличии в портрете смежных ключевых тем при малой актуальности в базе самой рубрики в нее попадает много низкорелевантных документов. Таким образом, все, что ложится на плечи пользователя при ручном редактировании инфопортрета, – это либо удаление омонимичных малозначимых для данной рубрики ключевых тем, либо увеличение ранга высокозначимых для данной рубрики ключевых тем. Кроме того, нужно выбрать походящее значение порога соответствия инфопортрета документа инфопортрету выборки, подходящее под конкретную задачу – подбора наиболее точных документов или же создания наиболее полной выборки, то есть порога соответствия точности и полноты.

Список литературы:

Антонов А.В. Методы классификации и технология Галактика-Zoom // сб. Международный форум по информации. М.: ВИНИТИ, 2003. Т.28

Антонов А.В., Козачук М.В., Мешков В.С. Галактика-Зум: Отчет об участии в семинаре РОМИП 2004. С. 133-141.http://romip.narod.ru/romip2004/10_koza4uk_Zoom.pdf

Антонов А.В., Курзинер Е.С. Автоматическое определение тематики большого необработанного текстового массива. // Материалы конференции «Диалог-2002». http://www.dialog-21.ru/archive_article.asp?param=7516&y=2002&vol=6078

Антонов А.В., Курзинер Е.С. Новые возможности поисково-аналитической системы «Галактика-ZOOM» (ранжирование документов по значимости). // Материалы конференции «Диалог-2003». http://www.dialog-21.ru/Archive/2003/Antonov.htm

Антонов А.В., Курзинер Е.С. Вычисление значимой части текста (в поисково-аналитической системе «Галактика-ZOOM»). // Материалы конференции «Диалог-2004». http://www.dialog-21.ru/Archive/2004/Kurziner.htm