МНОГОУРОВНЕВЫЙ КЛАССИФИКАТОР-НАВИГАТОР ПО ОТКЛИКАМ ИНФОРМАЦИОННО-ПОИСКОВОЙ СИСТЕМЫ

Д.В. Ландэ *(dwl@visti.ne)t,*

А.Н. Григорьев *(gri@visti.net)*

ИЦ «ЭЛВИСТИ», Киев, Украина

Описывается подход, модель и реализация построения многоуровневого классификатора-навигатора по откликам полнотекстовой информационно-поисковой системы. На основе определения близости слов построен интерфейс уточнения запросов, реализующий принцип «поисковых папок пользователя».

Общеизвестно, что большинство пользователей для поиска необходимой информации в Сети обращаются к известным поисковым системам, вводя в среднем 2-3 слова в качестве запроса. Очевидно, что найти требуемую информацию не так-то просто и они прерывают это занятие или пытаются перейти в специальные поисковые режимы [1]. Вместе с тем, хотя большинство современных поисковых систем и обладают необходимыми данными, вопрос визуализации, как правило, остается открытым.

Поэтому в последнее время получили распространение адаптивные интерфейсы уточнения запросов, чаще всего реализуемые путем кластеризации результатов первичного поиска. В работах авторов [2, 3] детально рассматривались такие подходы, как «информационные портреты» и методы "папок поиска" (). Существует множество систем, в которых реализованы подобные механизмы, например поисковые серверы Vivisimo, Mooter, iBoogie, которые предоставляют результаты поиска в виде кластеров – групп тематически подобных документов. На конференциях «Диалог» в свое время были представлены российская система «Галактика Зум», авторы представляли систему «InfoStream». В последнее время особых успехов в визуализации взаимосвязей информационных объектов добилась компания TouchGraph (www.touchgraph.com) с ее многочисленными продуктами типа GoogleBrowser, AmazonBrowser и WikiBrowser.

Данная работа посвящена не столько описанию алгоритмов кластеризации информации, многие из которых достаточно популярны, сколько обоснованию подхода к визуализации тематических кластеров, приближенной к традиционным статическим классификаторам. При этом ставится задача визуализации по отклику информационно-поисковой системы в режиме реального времени всего классификатора целиком, а не его отдельного фрагмента или уровня.

Опишем модель динамической классификации информации, которую можно рассматривать как некую «игру в слова», в некотором смысле напоминающую эволюционную игру Конвея. Именно игровой принцип позволил смоделировать классификатор-навигатор, который в результате нашел свое применение в реальном пользовательском интерфейсе. Правила игры, происходящей на плоскости, размеченной шестиугольными сотами, простые:

1 шаг: в соту вписывается слово, соответствующее некоторому понятию (в примере на рис. 1 – это слово «ураган»).

2 шаг: на основании анализа документального корпуса (представительного массива полнотекстовых документов) выбираются 6 наиболее связанных с первым словом значимых слов. Эти слова вписываются в соседние ячейки.

3 и последующие шаги: в свободные соты вокруг каждой из заполненных сот вписываются наиболее связанные с заполненными сотами слова (до 6, полученных из того же документального корпуса). При этом, если слова уже были использованы, то соседние соты остаются пустыми.

Процесс останавливается, когда добавление новых слов становится невозможным. На Рис. 1 процесс завершился за 4 шага.

Иллюстрация игры сотами вполне оправдана по двум причинам: с одной стороны, шестиугольники плотно покрывают плоскость, а с другой, количество вложений в классификаторе, не превышающее 6, соответствует принципам эргономики.

Понятно, что для информационного наполнения модели игры необходим достаточно мощный информационный ресурс. Такой ресурс был в распоряжении авторов изначально - это ретроспективная база данных системы контент-мониторинга InfoStream. Система InfoStream [3] применяется для решения задач автоматизированного сбора новостной информации с открытых web-сайтов и обеспечения доступа к ней в поисковых режимах. Эта разработанная в компании ElVisti система в настоящее время охватывает свыше 2000

источников, а ретроспективные базы данных системы представляют собой корпус объемом более 25 млн. документов.

Для построения классификатора-навигатора важнейшее значение имеет количество уровней - шагов игры. Что интересно, общая зависимость количества заполненных сот на каждом из шагов очень мало изменяется и слабо зависит от конкретного первого слова: практически максимум достигается на втором и третьем шагах, затем идет резкое угасание, хотя в некоторых случаях число «плодотворных» шагов достигает девяти-десяти.

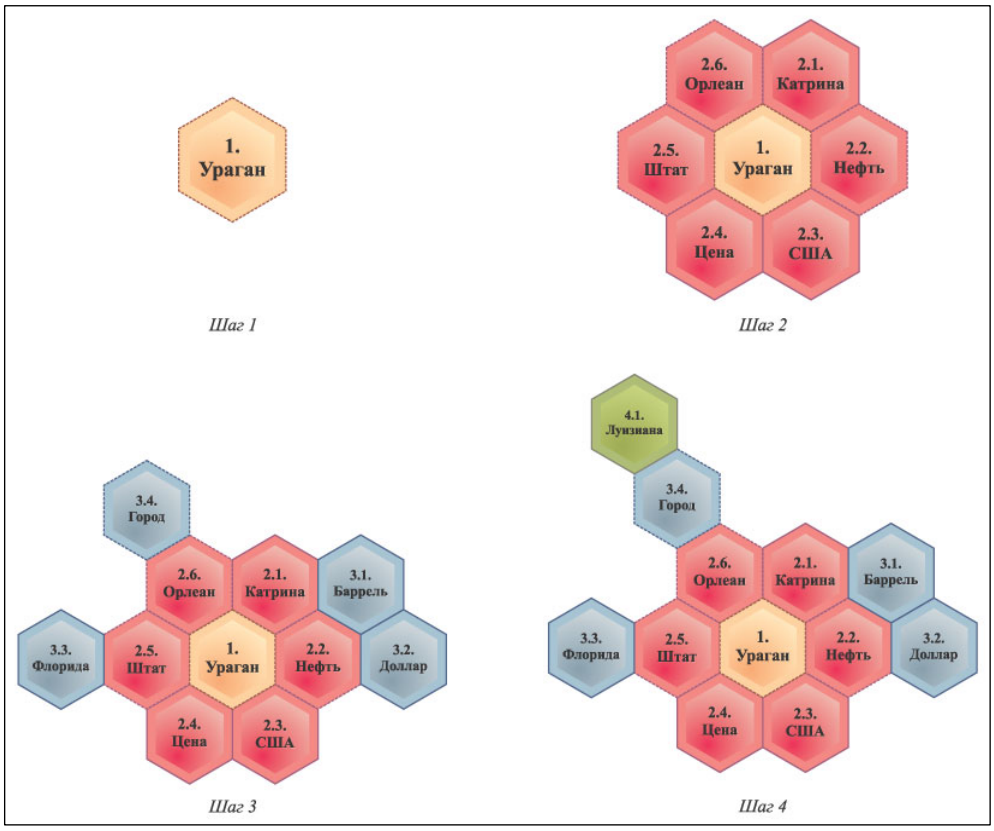

Рис. 1. Отражение в модели урагана “Катрина” (30 августа 2005 года)

Очевидно, что это количество зависит от двух параметров – объема первоначальной выборки (чем шире тематика, соответствующая слову, тем больше шагов требуется для завершения игры) и уровня «связности» документов из этой выборки (при этом связность можно понимать, например, как корреляцию массивов ключевых слов, входящих в отдельные документы). На Рис. 2 приведена полученная в результате многочисленных экспериментов сглаженная кривая средних значений количества слов (ось y) от номера шага (ось x). Названные выше два параметра позволяют аппроксимировать данную кривую формулой типа:

$y = \alpha f(x)g(x)$,

где α – некоторая нормирующая константа, $f(x)$ – медленно возрастающая функция, $g(x)$ – быстро убывающая функция, стремящаяся к нулю. Удивительно точным оказалось приближение:

$y = 14\,(x\text{-}1)^{\frac{1}{2}} e^{1-x}$.

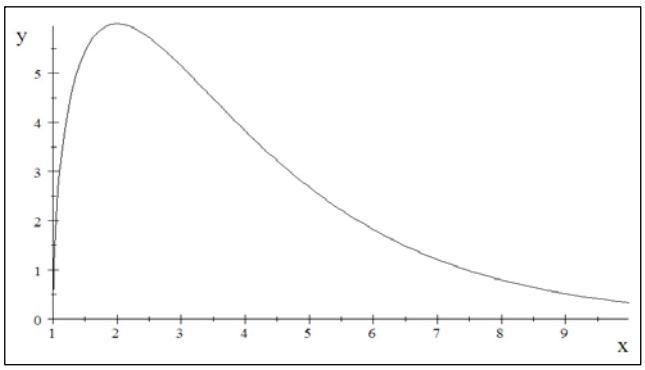

Рис.2. Среднее количество слов на каждом из шагов игры

Отдельный вопрос, касающийся модели, заключается в методике определения близости слов в документальном корпусе. В данном случае близость можно определить количеством документов, в которых слова появились совместно. Вместе с тем можно привести и более строгое правило, которое базируется на пространственно-векторной модели. Предположим, что W – это множество всех значимых слов, входящих в документальный корпус, $W = \{w_j\}$. Рассмотрим образ документа в пространстве слов из всего корпуса как вектор: $D_i = \{d^i_j\}$, где $d^i_j = 1$, если слово w_j есть в документе и 0 – в противном случае. Образом корпуса документов в этих обозначениях будем считать матрицу $D = \{D_i\}$, $i=1,..,N$, где N – количество документов в корпусе. В этом случае связность отдельных слов определяется матрицей D^TD, каждая ячейка которой определяет степень связности соответствующих ей слов.

Приведенная же выше игра стала основой для построения вполне серьезного инструмента – классификатора-навигатора в системе InfoStream. При этом были учтены такие особенности игровой модели:

Количество уровней (вложенность папок) не должно превышать трех. Этот вывод следует, с одной стороны, из анализа статистики модели, а, с другой, – из простейших требований юзабилити.

Для второго и третьего уровней количество близких слов может превышать шести, что позволит составить содержательно наполненные папки.

Правило отсутствия дублирования слов для второго и третьего уровней также целесообразно отменить.

В результате был предложен формальный алгоритм, реализующий навигацию в виде «папок поиска». На Рис. 3. представлен классификатор-навигатор по полнотекстовой базе данных, соответствующий первичному поисковому слову «бензин».

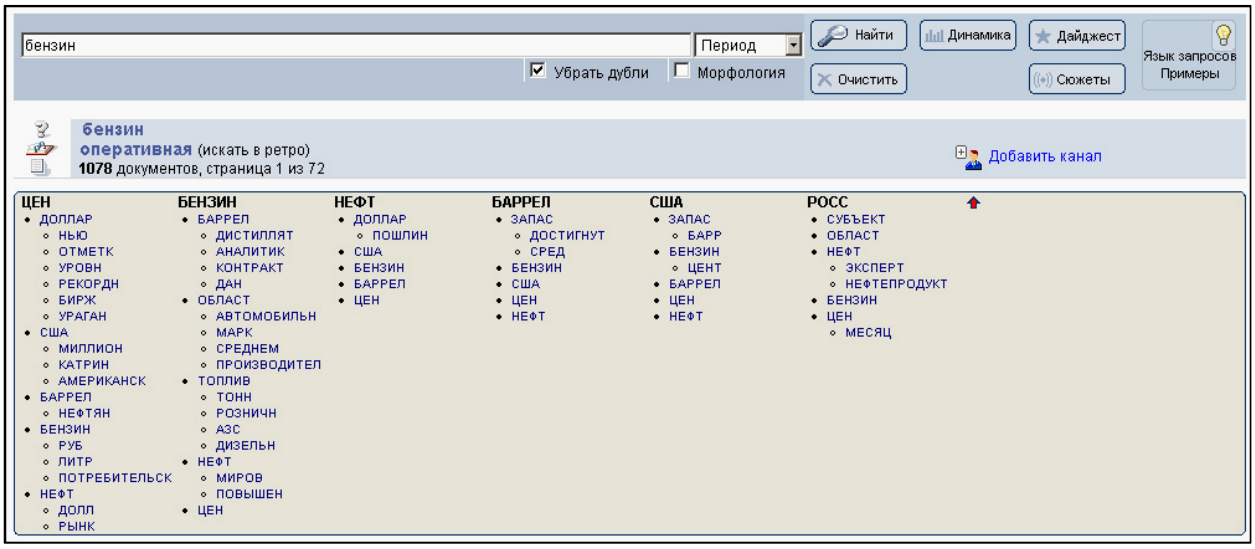

Рис.3. Классификатор-навигатор по слову “бензин”

Благодаря возможностям нового адаптивного инструмента, кластеризации результатов первичного поиска, система InfoStream позволила решать не только задачу поиска необходимой информации, но и, наряду с информационным портретом, возможностью построения сюжетных цепочек, дайджестов, решила задачу динамической визуализации, предоставила пользователям дружественный интерфейс для последующего аналитического обобщения данных.

Список литературы

1. Брайчевский С.М., Ландэ Д.В. *Современные информационные потоки: Актуальная проблематика // «Научно-техническая информация», серия 1, № 11. - 2005. - С. 21-33*

2. Григорьев А.Н., Ландэ Д.В. *Адаптивный интерфейс уточнения запросов к системе контент-мониторинга InfoStream//Труды Международного семинара «Диалог'2005». – 2005. – С. 109-111*

3. Ландэ Д.В. *Поиск знаний в Internet. Профессиональная работа. - М.: "Вильямс", 2005. - 272 с.*