ОСОБЕННОСТИ СООТНОШЕНИЯ ЛОКАЛЬНОЙ И ГЛОБАЛЬНОЙ ПОПУЛЯРНОСТИ СООБЩЕНИЙ ЭЛЕКТРОННЫХ СМИ

Снарский А.А., д.ф-м.н.,
Национальный технический университет «Киевский политехнический институт»
Ландэ Д.В., к.т.н., Брайчевский С.М., к.ф.-м.н., Григорьев А.Н., Дармохвал А.Т.,
Информационный центр «ЭЛВИСТИ»

Информационные потоки, генерируемые в Интернет обуславливают проблему навигации в сетевых ресурсах [1]. Одним из подходов к решению этой проблемы является ранжирование документов по уровню их популярности.

О популярности отдельного сообщения в информационном массиве можно говорить по тому, сколько в этом массиве имеется сообщений, подобных данному. Существует большое количество определений формального подобия, используемых в различных поисковых системах в режимах «поиска подобных документов». В частности, авторами использовался принцип определения подобия, применяемый в системе InfoStream [2]. Сообщение считается подобным исходному, если содержит определенное количество наиболее значимых слов из него (назовем этот критерий α-подобием). Принцип выявления значимых слов и их количесвта базируется как на статистическом алгоритме, построенном на основе закономерности Ципфа, так и на некоторых эмпирико-лингвистических подходах. Этот критерий успешно применяется в течение нескольких лет.

Под глобальной популярностью для каждого сообщения понимается количество α-подобных ему сообщений в ретроспективной базе данных (свыше 1 млн. документов по выбранной теме за последний год). Под локальной популярностью понимается количество α-подобных сообщений за тот день, когда появилось исходное сообщение.

Исследуемый тематический массив документов за последние 3 дня объемом около 5000 сообщений был ранжирован по глобальной популярности. Был построен соответствующий график, близкий по форме к гиперболе (ось X – номер документа, в соответствии с рангом, ось Y – популярность). Для каждого из сообщений, ранжированных указанным выше образом был также построен график локальной популярности. Для различных участков второго графика было доказано свойство скейлинга (самоподобия), что является необходимым условием фрактальности [3].

Наряду с этим результатом, было выявлено некоторое количество сообщений, характеризующихся большим соотношением локальной популярности к глобальной. Этот факт позволяет судить о событиях, описываемых в данных сообщениях, как о новых. Таким образом получен алгоритм выявления документов, получивших большую популярность только в последнее время, что является частным решением актуальной научно-практической проблемы выявления новых событий (New Event Detection) [4].

Литература

1. Ландэ Д.В. Основы интеграции информационных потоков - К.: Інжиніринг, 2006. - 240 с.
2. Григорьев А.Н., Ландэ Д.В., Бороденков С.А., Мазуркевич Р.В., Пацьора В.Н. InfoStream. Мониторинг новостей из Интернет: технология, система, сервис: научно-методическое пособие. – К.: Старт-98, 2007. – 40 с.
3. Федер Е. Фракталы / -М.: Мир, 1991, -254 с.
4. Ландэ Д.В., Фурашев В.Н. Выявление новых событий в рамках системы контент-мониторинга // Научно-техническая информация.Сер.2.Информационные процессы и системы №12 – 2006. - С. 17-20.