

Доклад, посвященный ранжированию источников новостей, на семинаре "Поисковые технологии 2010?"

26-28 февраля 2010 г. компания "Ашманов и Партнеры" собрала разработчиков поисковых систем на выездном семинаре "Поисковые технологии 2010?" в спортивно-развлекательном парке "Яхрома". Специалисты компаний ("Рамблер", "Яндекс", "Нигма", "Ашманов и Партнеры", Meta.ua, "Галактика", ELVisti и др.) обсудили проблемы релевантности и индексации, ситуацию с поисковым спамом и будущее поисковых систем.

Семинар был посвящен проблемам разработки и развития поисковых проектов и технологий и рассчитан на менеджеров и разработчиков поисковых систем и поисковых сервисов, прикладных лингвистов, специалистов по поиску, студентов и научных сотрудников профильных вузов и кафедр. Пресса и SEO-специалисты специально не приглашались.

На семинаре с совместным докладом от компаний "Галактика" и ELVisti по теме "Рейтинг онлайн-СМИ на основе дублирования новостей" выступил А. Антонов.

Ранжирование результатов выдачи поисковых систем, в частности новостных ресурсов, является одной из главных задач, стоящих перед современными поисковыми технологиями. В докладе были приведены методы создания рейтинга онлайн-СМИ для агрегирующего новостного ресурса www.webground.su.

Ранжирование источников основано на информации о группах найденных новостей-дубликатов и признаком времени публикации, приписанном новостям. На первом этапе алгоритм поиска дублирующихся сообщений разбивает множество новостных сообщений на непересекающиеся подмножества. После чего в каждом подмножестве сообщения ранжируются по времени публикации в убывающем порядке. Каждое из выделенных подмножеств представляется в виде направленного графа, вершинами которого являются сообщения, а ребрами - отношения в упорядочении внутри подмножества.

С целью сокращения вычислительной сложности алгоритма принято ограничение, при котором ребра могут соединять только соседние элементы упорядочения. Каждое из ребер направлено от более раннего к более позднему сообщению. К построенным графам применен алгоритм PageRank, с помощью которого каждой из вершин-сообщений на графе присваивается соответствующий вес. Использована версия алгоритма PageRank с ненормированными весами ребер. Для составления итогового рейтинга источника учитывается накопленная месячная статистика веса PageRank сообщений источника и среднее время запаздывания при публикации новостей.

Анализ результатов формирования рейтинга (пример:
<http://webground.su/sources.php?param=SourceList&Sort=2&Filter=0>).

Презентация в формате PDF:
<http://dwl.visti.net/art/yahroma/yahroma.pdf>

Инф. ИЦ "ЭЛВИСТИ"