Андреев А.Ю.
УНК "ИПСА" НТУУ "КПИ"

Автоматизированное средство поиска, систематизации и мониторинга информации в интернете – SiteSputnik

Непрерывно увеличивающиеся объемы доступной в Интернет информации, в том числе оперативной, делает проблему поиска необходимых сведений весьма актуальной и сложной. Для автоматизации этой задачи разработаны различные, как зарубежные, так и отечественные системы поиска, представляющие собой Web-страницы специального вида. Однако, несмотря на наличие многочисленных средств автоматизации поиска, эта задача остается достаточно трудоемкой, требующей от пользователя определенного опыта, интуиции, глубокого знания терминологии предметной области. Также она усложняется тем, что на данный момент большинство поисковых машин являются конкурентами и не используют поисковые базы совместно, используют разные языки запросов, по разному индексируют страницы и оценивают их релевантность. Основными характеристиками поиска являются полнота и точность.

Изучая некую тему перед пользователем часто встает задача регулярного мониторинга доступной в интернете информации по одним и тем же запросам, конечно же объемы информации найденной по нескольким ключевым словам обычно слишком велики для обработки, а сложные запросы не обрабатываются онлайновыми средствами поиска достаточно эффективно. Аналогами представленного в этой работе решения было воспользоваться сложно по причине того, что найденная информация нуждалась в личной обработке, особенно в случае обновления уже проанализированных источников. Практически все доступные поисковые средства не предусматривают такой возможности.

С целью автоматизации решения этой задачи используются клиентские приложения, которые являются посредниками между пользователем и поисковыми машинами. Одним из таких приложений является SiteSputnik [2]. Это программное обеспечение обладает своим синтаксисом для ввода запроса, набранный запрос так же дублируется в поисковые машины указанные пользователем, но найденные источники и тексты обрабатываются таким образом, чтобы результаты поиска не повторялись, веб-страницы оцениваются программой по нескольким критериям (позиции на которых эти ресурсы были выданы каждым поисковиком, частота встречаемости ключевых слов и т.д.) с целью увеличения точности. Результаты поиска заносятся в базу данных и при повторном поиске осуществляется проверка на совпадения. Пользователю выдаются страницы, помечаются те, которые не были найдены ранее, а так же документы и участки текста в них которые были изменены с момента последнего обнаружения. Классические инструменты поиска (онлайн каталоги, поисковые машины и метапоисковые машины) эту задачу не решают, что создает нишу для реализации нового класса ПО – посредника между поисковыми системами и пользователей.

Список литературы

1. Мыльников А.Б. Список публикаций по программе. Режим доступа: http://ab.vlink.ru/t_frame.htm