ИСПОЛЬЗОВАНИЕ РУССКОГО ЯЗЫКА В СИСТЕМАХ ИНФОРМАЦИОННОГО ПОИСКА И АНАЛИЗА

Рустамов Заур Низами оглы

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Петрова Н.Е. – к.филол.н., доцент

Статья посвящена анализу лексико-семантических особенностей ключевых слов, функционирующих в информационно-поисковых системах Интернета. Рассматривается их частотность в популярных поисковых системах, система принципов номинации, семантические и ассоциативные связи ключевых слов в русскоязычном Интернете.

Функционирование русского языка в Интернете вызывает значительный интерес исследователей. Важную роль в оценке эффективности текстов, представленных в Интернете, играет такой показатель, как ключевые слова, т. е. слова, которые используются при поиске определённой информации. Понятие «ключевые слова» изначально используется в литературоведении, где под ними понимают единицы лексического уровня языка, которые обозначают узловые понятия в тексте и актуализируют концепт, являются фиксаторами наиболее важной для читателя информации, необходимой для понимания смысла [1, с. 87]. Используется понятие ключевых слов и в сфере информационных технологий. Там оно обозначает запросы, по которым тот или иной сайт может быть найден в поисковых системах.

Поисковыми системами Интернета проводится статистика запросов пользователей с помощью ключевых слов. В большинстве случаев при работе с сервисом статистики имеется возможность классифицировать результаты поиска по датам, географии запросов. При этом, как правило, сервис показывает не только данные об искомом запросе, но также и о словосочетаниях, синонимах и близких темах. Например, если сайт посвящен туризму и отдыху, то он должен появляться в системах поиска (например, Яндекс, Google) на основе определённого лексического запроса: «отдых», «курорт», «горящие путевки», названия туристических фирм и т.п.

Ключевые слова несут важную информацию об интересах аудитории, в том числе дают представление о национально-ориентированной специфике актуальности электронных текстов на том или ином языке. Ключевые слова косвенным образом характеризуют и саму аудиторию, например, в 2006 г. наиболее частотными ключевыми словами российского Интернета были такие лексемы, как «реферат», «рефераты», «банк рефератов», «гороскоп», «погода», и это даёт представление о том, что основными пользователями Интернета являются молодые люди в возрасте от 18 до 32 лет, в большинстве студенты, которых часто интересует информация образовательного характера [2].

Система формирования ключевых слов является составной частью так называемой контекстной рекламы, которая представляет собой информацию рекламного содержания. Отличительной чертой контекстной рекламы является то, что она появляется там, где с помощью ключевых слов ищут ту или иную информацию. Контекстная реклама возникает в связи с конкретным тематическим запросом в системе поиска с помощью ключевых слов. Например, если вводится словосочетание «русский язык», то контекстная реклама указывает коммерческие курсы, которые предлагают обучение русскому языку. Если в систему поиска вводится ключевое слово «школа», то в контекстной рекламе появляются сообщения об адресах школ, о

сайтах школ, справочниках. Поиск по ключевым словам оказывает огромное влияние на популярность того или иного сайта.

Необходимо подчеркнуть, что отмечаются изменения в актуальности тех или иных ключевых слов в зависимости от внешних событий или обстоятельств. Яркий пример изменения в количестве ключевых слов во время террористической атаки 11 сентября в США или чемпионата Европы [3] по футболу. Следует отметить, что, несмотря на повышение актуальности тех или иных слов в связи с определенными событиями, общая тенденция выбора ключевых слов для данной лингвокультурной общности остается стабильной.

В анализе запросов Яндекса были опубликованы результаты анализа лингвистических особенностей ключевых слов. Около 2,5% запросов сформулированы как вопрос. Это обычные вопросы, и пользователи, которые их задают, общаются с поисковой системой так, как будто это живой человек. К примеру, у Яндекса спрашивают «Как пройти в библиотеку?» в среднем 41 раз в месяц, Зачем Герасим утопил Муму? – 53 раза, Кто убил Лору Палмер? – 107 раз. Вопросов, начинающихся со слова «как», задают больше, чем вопросов, начинающихся со всех остальных вопросительных слов. Интересно, что вопрос «что?» значительно популярнее вопроса «кто?», и в частности вопрос «что делать?» интересует пользователей больше вопроса «кто виноват?» В настоящее время автокоррекция текста стала неотъемлемой частью многих приложений и операционных систем. Она помогает пользователям писать текст быстрее и более точно, предлагая исправления на лету [4].

Самые короткие составляющие запросов к Яндексу – отдельные буквы и цифры. Пользователей интересуют все буквы русского алфавита, больше всего – предлоги и однобуквенные союзы (например, «и» в августе 2008 г. искали 4 385 966 раз), а меньше всего букву «ъ» (9 тысяч запросов в месяц). Наиболее длинные осмысленные слова, как правило, сложные существительные, состоящие из нескольких корней. Самое длинное слово, заданное в качестве запроса к Яндексу в августе 2008 г., состоит из 37 символов «гиппопотомомонстросесквиппедалиофобия». По этому запросу находится 4583 страницы (на сентябрь 2008 г.). Среди самых длинных запросов, на которые существуют ответы в Интернете, преобладают различные химические соединения (этилоксиэтилпарафенилендиаминсульфат, 35 символов), названия компаний («Средневолжскэльэлектросетьстрой», 32 символа) и разного рода фобии (например, гексакосиойгексеконтагексофобия, 31 символ) [5].

В запросах на поиск картинок самые длинные запросы это «электростеклоподъёмник» и «электроводонагреватель», по 22 символа. Для общения с поисковой машиной чаще всего используют существительные, эту часть речи содержат 75% запросов к поиску. Вторая по распространённости часть речи – прилагательные, они присутствуют в 16% запросов к веб-поиску. Глаголы используют только в 5% случаев, а наречия отмечены менее чем в 1% [6].

Существующие системы анализа ключевых слов позволяют эффективно использовать различные инструменты, показатели для поиска и выбора ключевых слов. В основе сбора данных по ключевым словам лежат два принципа – принцип счётчика (специального скрипта, загружающегося пользователю сервером вместе с загрузкой веб-страницы) и принцип анализа логов (специальных файлов на сервере, фиксирующих все посещения). Оба способа сбора информации работают независимо друг от друга и каждый с определенной степенью погрешности. Из собранной с помощью счетчика или лог-анализатора информации можно формировать разные массивы данных, изучать отдельные срезы и тематические выборки. Подобные обобщенные данные часто можно встретить в исследованиях по глобальной статистике Рунета, например, HotLog и SpyLog.

Мы считаем, что предназначение любого интернет-ресурса в первую очередь заключается в эффективном достижении целей, определённых при его создании. Эффективность сайта зависит от его содержания, определяется числом посетителей сайта (в частности, приходящих из поисковых серверов), скоростью и удобством получения интересующей информации, количеством повторных возвращений на данный сайт. Эффективность сайта также зависит от того, насколько интересны статьи, удобна навигация, привлекателен дизайн и т.п. Для оценки эффективности сайта собираются статистические данные посещения данного ресурса, применяются различные методы математической статистики для получения и обработки результатов. Интернет-статистика даёт возможности анализировать различные особенности сайта. Во-первых, можно существенно улучшить дизайн, навигацию и размещение ссылок на сайте. Во-вторых, на основании полученных статистических отчётов возможно повысить рейтинг сайта на крупнейших поисковых системах Рунета (Яндекс, Google, Rambler) [7].

Таким образом, предварительный анализ показывает, что ключевые слова вступают в Интернете друг с другом в особые отношения, в первую очередь на основе гипертекстовых связей, тематического единства, разного рода ассоциаций. Можно увидеть, что иногда возникают непривычные с точки зрения классической лексикологии семантические связи слов. Например, к ключевому слову «виза» интернет-системой будут предложены названия фирм, адреса посольств, путеводители, карты дорог, адреса гостиниц и др. В свою очередь отметим, что ключевые слова, используемые в поисковых системах Интернета, очень важны, они выполняют различные функции: помогают найти текст с соответствующим содержанием; повышают посещаемость (тем самым и рейтинг того или иного сайта), если в текст включаются наиболее частотные или востребованные в данный момент ключевые слова.

Список использованных источников:

1. Пятрова, Н. Я. Ключавыя словы ў ідыялекце Міхася Зарэцкага / Н. Я. Пятрова // Беларуская лінгвістыка. – 2014. – Вып. 72. – С. 87 – 94.

2. Лингвистический дизайн WEB-страниц [Электронный ресурс]. – Режим доступа: http://lib-repository.mephi.ru/conferences_mephi/2018_MATEMATIKA_I_MATEMATICHESKOE_MODELIROVANIE_Sbornik.pdf#page=117. – Дата доступа: 27.03.2024.

3. Язык Интернет-коммуникации [Электронный ресурс]. – Режим доступа: https://cyberleninka.ru/article/n/autocorrection. – Дата доступа: 27.03.2024

4. Разработка интеллектуальной функции автокоррекции и исследование обучающей способности алгоритмов распознавания естественного языка [Электронный ресурс]. – Режим доступа: https://www.elibrary.ru/item.asp?id=30744162. – Дата доступа: 27.03.2024.

5. . Языковой вкус интернет-эпохи в России: функционирование русского языка в Интернете: концептуально-сущностные доминанты [Электронный ресурс]. – Режим доступа: http://ucom.ru/doc/na.2016.03.03.083.pdf. – Дата доступа: 27.03.2024.

6. Функция анализа как технология обработки данных [Электронный ресурс]. – Режим доступа: https://cyberleninka.ru/article/n/autocorrection-pomoschnik-kak-tehnologiya-obrabotki-dannyh. – Дата доступа: 27.03.2024.

7. Искусственный интеллект в анализе интернета. Пути развития [Электронный ресурс]. – Режим доступа: http://lib-repository.mephi.ru/conferences_mephi/2018 _Sbornik.pdf#page=117. – Дата доступа: 27.03.2024.