СРАВНИТЕЛЬНЫЙ АНАЛИЗ ЛИНГВИСТИЧЕСКОЙ ЧАСТИ ПОПУЛЯРНЫХ РУССКИХ И ВЬЕТНАМСКИХ ПОИСКОВЫХ СИСТЕМ

Нгуен Куок Дай, Фам Куанг Биен

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

Петрова Н.Е. – к.филол.н., доцент

В статье рассматриваются основные характеристики, достоинства и недостатки поисковых систем. Анализируются некоторые популярные вьетнамские и русские поисковые системы.

Поисковая система – это компьютерная система, предназначенная для поиска информации. Одно из наиболее известных применений поисковых систем – веб-сервисы для поиска текстовой или графической информации во Всемирной паутине. Существуют также системы, способные искать файлы на FTP-серверах, товары в интернет-магазинах, информацию в группах новостей Usenet [1].

Самые популярные поисковых систем в мире: Гугл, Яндекс, Сос Сос. Согласно данным за сентябрь 2019 года, доля рынка Гугл в мире поиска составляет 69,29%, Яндекс – 53,3% (в России), а Сос Сос – 25% (во Вьетнаме).

Для поиска информации с помощью поисковой системы пользователь

формулирует поисковый запрос. Работа поисковой системы заключается в том, чтобы по запросу пользователя найти документы, содержащие либо указанные ключевые слова, либо слова, как-либо связанные с ключевыми словами. При этом поисковая система генерирует страницу результатов поиска. Такая поисковая выдача может содержать различные типы результатов, например: веб-страницы, изображения, аудиофайлы. Некоторые поисковые системы также извлекают информацию из подходящих баз данных и каталогов ресурсов в Интернете. Поисковая система тем лучше, чем больше документов, релевантных запросу пользователя, она будет возвращать. Результаты поиска могут становиться менее релевантными из-за особенностей алгоритмов или вследствие человеческого фактора [2].

Основные составляющие поисковой системы: поисковый робот, индексатор, поисковик. Как правило, системы работают поэтапно. Сначала поисковый робот получает контент, затем индексатор генерирует доступный для поиска индекс, и наконец, поисковик обеспечивает функциональность для поиска индексируемых данных. Чтобы обновить поисковую систему, этот цикл индексации выполняется повторно [3].

Немного в стороне от статистических моделей и структур данных стоит класс алгоритмов, которые традиционно относят к лингвистическим. Точно границы между статистическими и лингвистическими методами провести трудно. Условно можно считать лингвистическими те методы, которые опираются на словари (морфологические, синтаксические, семантические), созданные человеком. Хотя считается доказанным, что для некоторых языков (например, для английского) лингвистические алгоритмы не вносят существенного прироста точности и полноты, все же основная масса языков требует хотя бы минимального уровня лингвистической обработки. Не вдаваясь в подробности, приведем только список задач, решаемых лингвистическими или окололингвистическими приемами:

– автоматическое определение языка документа;

– токенизация (графематический анализ): выделение слов, границ предложений, исключение неинформативных слов (стоп-слов);

– лемматизация (нормализация, стемминг): приведение словоизменительных форм к «словарной», в том числе и для слов, не входящих в словарь системы;

– разделение сложных слов (компаундов) для некоторых языков (например, немецкого);

– дизамбигуация: полное или частичное снятие омонимии, выделение именных групп [4].

Сравнивая поисковые системы во вьетнамском и русском языках, мы можем увидеть как совпадения, так и расхождения в них. Понаблюдаем, в каких системах поисковый запрос выполняется лучше. На рисунках 1 и 2 мы привели примеры работы такой поисковой системы, как Гугл:

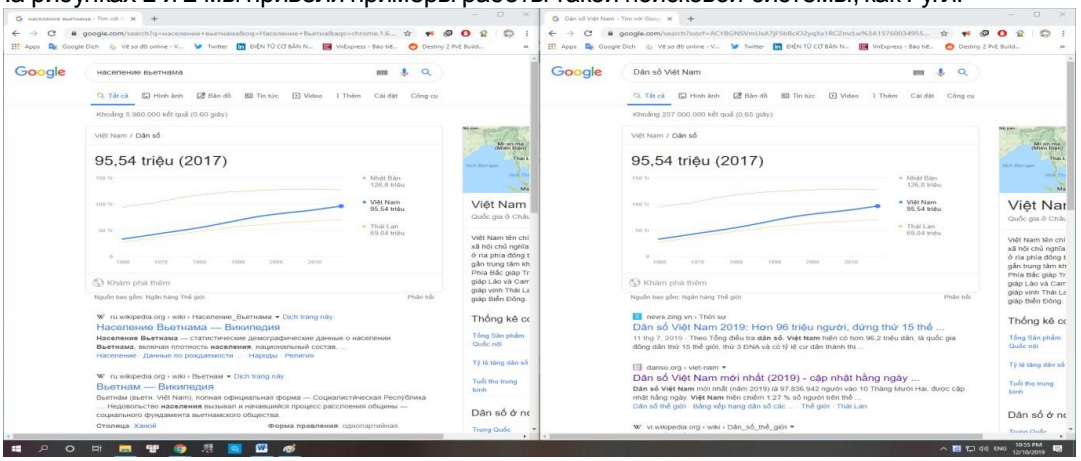

Рисунок 1 – Запрос «*население Вьетнама*»

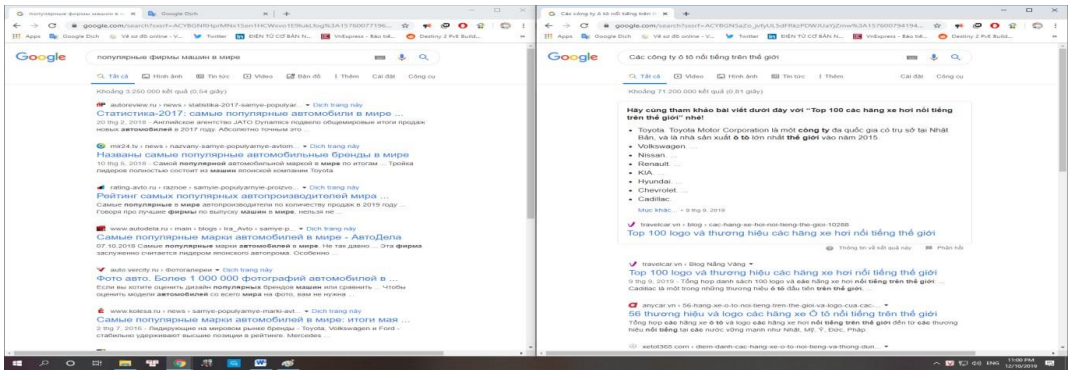

Рисунок 2 – Запрос «*популярные фирмы машин в мире*»

Мы ввели в Гугл одинаковые запросы на русском и на вьетнамском языках. Первый запрос: *население Вьетнама* (рус.) – *Dân s Việt Nam* (вьетн.). Второй запрос: *популярные фирмы машин в мире* (рус.) – *C c công ty ô tô n i tiếng trên thế gi i* (вьетн.).

Разница между результатами поиска отчетливо видна, несмотря на то, что цель поиска совпадает. Когда ключевое слово на русском и на вьетнамском языках для поиска простое, мы видим, что результаты поиска очень похожи. Когда ключевое слово для поиска является более сложным, как во втором запросе, результаты поиска менее схожи. Так, поиск в первом случае одинаковый, а во втором случае на вьетнамском языке Гугл сразу показал самые популярные фирмы машин.

Далее рассмотрим несколько примеров использования таких поисковых систем на вьетнамском и русском языках, как Яндекс и Coc Coc (рисунок 3):

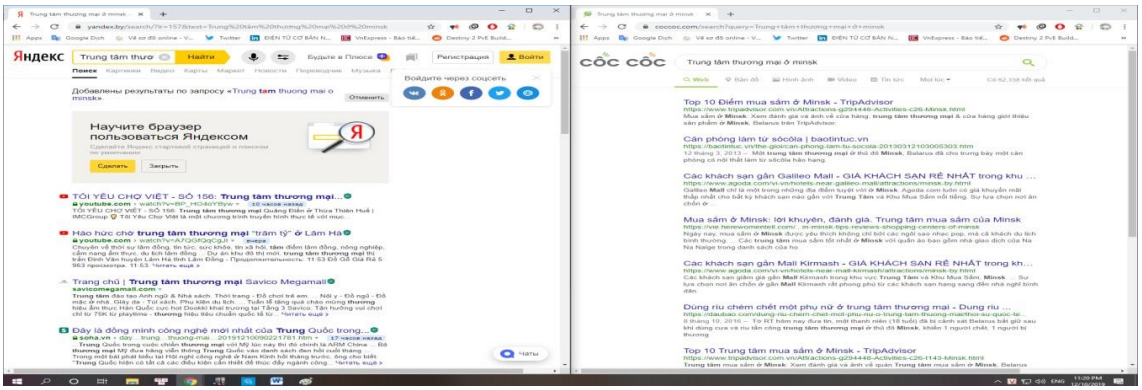

Рисунок 3 – Выполнение поиска в Яндекс и CocCoc

На следующем рисунке мы показали выполнение одинакового поискового запроса «*торговый центр в Минске*» в Яндекс и Гугл (рисунок 4):

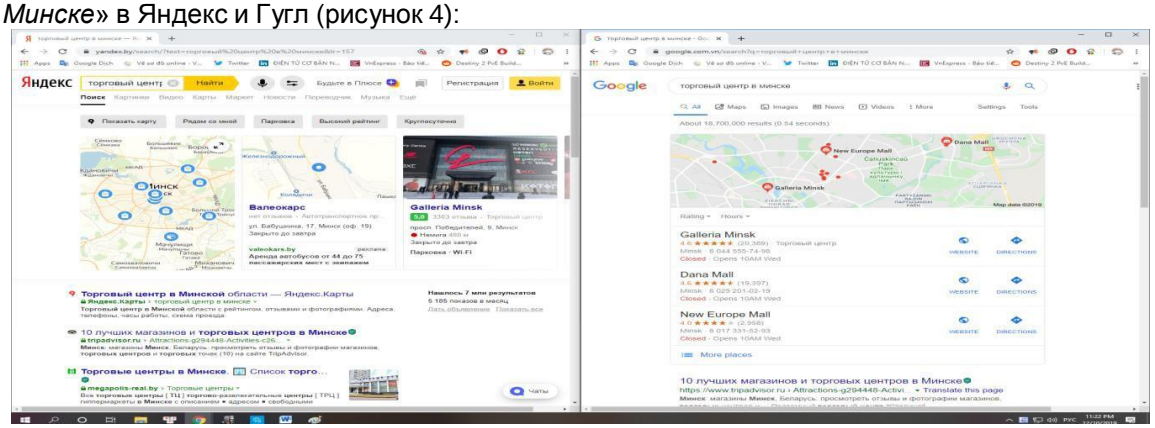

Рисунок 4 – Выполнение поиска в Яндекс и Гугл

На рисунке 5 мы видим выполнение поисковых запросов на вьетнамском языке (рисунок 5):

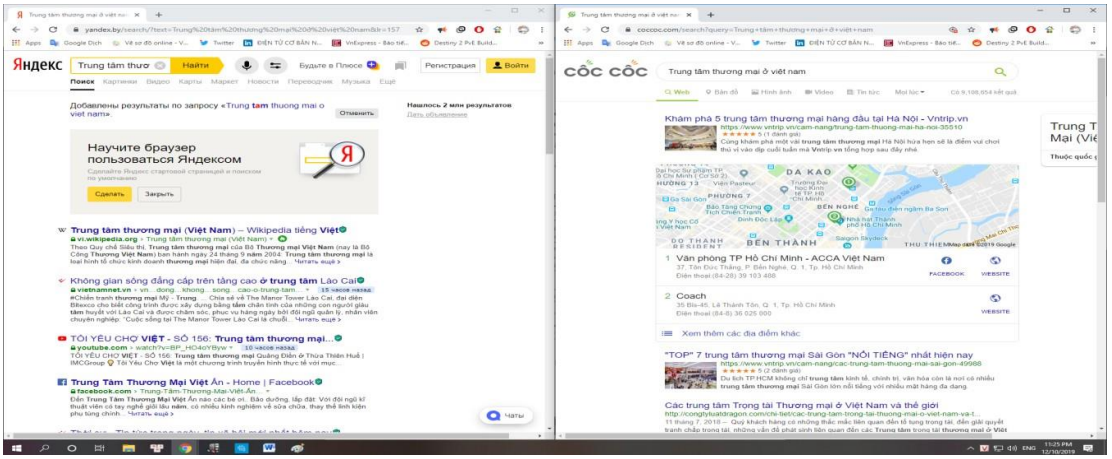

Рисунок 5 – Выполнение поиска на вьетнамском языке

Из результатов поиска видно, что поисковая система Яндекс поддерживает ключевые слова на русском языке, а Coc Coc не поддерживает, Coc Coc поддерживает вьетнамский язык и показывает более точные результаты на вьетнамском языке, чем Яндекс. В случае, когда ключевое слово для поиска связано с информацией о Вьетнаме, поисковая система Coc Coc работает более эффективно.

Таким образом, вьетнамская поисковая система Coc Coc более эффективна для вьетнамцев, а русские поисковые системы Гугл и Яндекс более точны для русскоязычных запросов.

Сделаем вывод. В мире существует множество разных поисковых систем на разных языках. В каждой поисковой системе лингвистика играет важную и незаменимую роль. Поэтому необходимо пользоваться системой, которая адаптирована под родной язык пользователя.

Список использованных источников:

1. Поисковая система [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org/wiki/Поисковая_система. – Дата доступа: 04.12.2019.

2. Поисковая система [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org/wiki/Поисковая_система. – Дата доступа: 04.12.2019.

3. Поисковая система [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org/wiki/Поисковая_система. – Дата доступа: 04.12.2019.

4. Как работают поисковые системы [Электронный ресурс]. – Режим доступа: https://habr.com/ru/company/yandex/blog/464375/.– Дата доступа: 27.08.2019.