

ИСПОЛЬЗОВАНИЕ МОДУЛЯ МОРФОЛОГИЧЕСКОГО АНАЛИЗА В LUCENE.NET

М.В. ЖЕРДЕВА, *асп. ГБОУ ВО МО «Технологический университет»⁽¹⁾*,

В.М. АРТЮШЕНКО, *проф, ГБОУ ВО МО «Технологический университет», д-р техн наук⁽¹⁾*

masha8908@rambler.ru

⁽¹⁾ ГБОУ ВО МО «Технологический университет», 141070 Московская область, г. Королев, ул. Гагарина, д. 42

Рассматривается использование морфологического анализа в Lucene.Net. Актуальность данной статьи не вызывает сомнения, т.к. разработчик, реализующий какой-либо проект, сталкивается с потребностью реализовать поиск в своем веб-приложении. Автор рассматривает понятия стемминга и лемматизации. Выделена цель стемминга и лемматизации. Показано применение модуля морфологического анализа в Lucene.Net. Lucene.Net – это перенесенный с платформы Java поисковый движок в lucene. В lucene – это высокопроизводительная, масштабируемая библиотека для полнотекстового поиска. Полнотекстовый поиск документов относится к процессу поиска, информации в документах или метаданных о документах. В lucene позволяет добавлять возможности поиска в различные приложения. Главной особенностью библиотеки является то, что требуется малый объем памяти, наличие ранжированного поиска, возможность одновременного поиска и обновления индекса. Введение в настоящее время и на протяжении уже несколько лет является самой популярной свободной библиотекой полнотекстового поиска. Рассмотрены цели, которые нужно достичь при реализации операций, выполняемых модулем морфологического анализа. Предложена идея модификации модуля лемматизации Lucene.Net для быстрого и релевантного поиска ключевых слов.

Ключевые слова: полнотекстовый поиск, стемминг, лемматизация, лемма, стоп-слова, токен.

Lucene.Net – это перенесенный с платформы Java поисковый движок Lucene. Он поддерживает тот же API и те же классы, что и оригинальная версия. Это накладывает определенный отпечаток, а также делает индекс обратно совместимым для обеих платформ [17]. Стабильный релиз 3.0.3 от 26 октября 2012 г. может быть загружен из официально-го репозитория.

Lucene – это высокопроизводительная, масштабируемая библиотека для информационного поиска (ИП). ИП относится к процессу поиска документов, информации в документах или метаданных о документах. Lucene позволяет добавлять возможности поиска в различные приложения. Этот открытый проект реализован Apache Software Foundation и может быть использован на условиях лицензии Apache Software. Главной особенностью библиотеки является то, что требуется малый объем памяти, наличие ранжированного поиска, возможность одновременного поиска и обновления индекса, поиск, основанный на «полях» и т. д. Для другого движка нет такого количества портов на другие языки, как для Lucene. Одним из факторов такой популярности является то, что можно строить решения, работающие с индексом и производящие поиск, но не имеющие собственного индексатора.

Помимо основных модулей Lucene, существует большое количество дополнений, предлагающих полезные расширения функциональности. Некоторые из них являются жизненно необходимыми почти для всех приложений, например модуль проверки орфографии и подсветки результатов. используется в таких проектах, как StackOverflow, Twitter Loggly.com, Archive.org, Search.USA.gov WhiteHouse.gov, Digg, MySpace LinkedIn, Fedex, Apple SalesForce.com, Encyclopedia Britannica CD-ROM/DVD, Eclipse IDE Atlassian (JIRA) и др. [3, 4, 8–10].

Таким образом, Lucene в настоящее время и на протяжении уже несколько лет является самой популярной свободной библиотекой для ИП.

Одним из полезных средств, используемых в механизмах полнотекстового поиска, является стемминг.

Стеммингом обычно называется приближенный эвристический процесс, в ходе которого от слов отбрасываются окончания в расчете на то, что в большинстве случаев это себя оправдывает. Стемминг основан на правилах морфологии языка и не требует хранения словаря всех слов. [6]

Таким образом, стемминг отсекает от слова окончания и суффиксы, чтобы остав-

шаяся часть была одинаковой для всех грамматических форм слова.

Более сложным подходом к решению проблемы определения основы слова является лемматизация.

Лемматизацией называется преобразование слова в словарный вид или лемму. Данный метод используется в алгоритмах поисковиков при индексировании интернет-страниц. Процесс дает возможность хранения данных страницы набором слов в индексе для удобной схематизации файлов. Это позволяет ускорить индексацию и сформировать более четкий ответ на поисковый запрос, так как сокращенную форму слова поисковик анализирует быстрее.

Лемма – это первоначальная, основная форма слова. Для существительных и прилагательных ею является форма единственного числа, именительного падежа. Для глаголов лемма является инфинитивом, неопределенной формой слова, отвечающей на вопрос в инфинитиве [4–6].

Лемматизатором называют программу или модуль программы, проводящий лемматизацию. Главной его задачей является то, что он улучшает релевантность поиска. Также лемматизатор уменьшает количество анализируемых слов.

Цель стемминга и лемматизации одна – привести словоформы и производные формы слова к общей основной форме.

В Lucene.Net в механизме полнотекстового поиска используется стемминг. Механизм стемминга основан на использовании алгоритма Портера.

Основная идея алгоритма Портера заключается в том, что существует ограниченное количество формо- и словообразующих суффиксов и основа слова преобразуется без использования каких-либо баз (словарей) основ: только множество существующих суффиксов (при этом сложные составные суффиксы разбиваются на простые) и вручную заданные правила. В алгоритме Портера была описана последовательность шагов, в каждом из которых при определенных правилах может происходить одно из определенных преобразований окончаний.

То, что алгоритм Портера не использует никаких словарей и баз основ, является плюсом для быстрого действия и спектра применения (он неплохо справляется с несуществующими словами) и одновременно минусом с точки зрения точности выделения основы. Кроме того, к минусам алгоритма Портера часто относят человеческий фактор: то, что правила для проверки задаются вручную и иногда связаны с грамматическими особенностями языка, увеличивает вероятность ошибки.

Модуль морфологического анализа в Lucene.Net

В Lucene модуль морфологического анализа обычно выполняет преобразование множества всех слов языка во множество лемм – нормализованных форм слов. В литературе данный модуль поисковой системы называют модулем морфологического анализа, нормализатором слов, лемматизатором.

Операцию, выполняемую данным модулем можно представить, как отображение: $T \rightarrow L$, где T – множество всех терминов; L – множество всех лемм, при этом количество лемм меньше мощности множества всех терминов $|T| > |L|$.

Реализуя данное преобразование нужно стараться достичь следующих целей:

1. Увеличение полноты поиска: отбираются документы, которые содержат все формы слова, поэтому в результат поиска попадают не только документы со словом в совпадающей с запросом форме, но и другие документы, содержащие различные формы данного слова.

2. Увеличение точности поиска: при использовании статистических алгоритмов поиска и отбора в результат поиска нескольких документов, которые получили наибольший вес, очень важным становится получение частотных характеристик документов. Использование вместо частот слов частоты лемм может позволить получить больший вес для релевантных документов и тем самым поместить их во множество отобранных.

3. Уменьшение размера индексной информации и ускорение обработки запроса.

Лемматизация приводит к уменьшению размера индекса и увеличению скорости обработки запроса, т.к. количество лемм меньше количества слов. Развитие методов сжатия индексной информации и совершенствование алгоритмов поиска приводит к тому, что использование морфологической обработки для уменьшения размера индекса в настоящее время не является критически важным в большинстве случаев.

Заключение

Высокая скорость является критерием эффективного индексирования. Она зависит от количества форм слова — чем их меньше, тем раньше закончится схематизация документа.

Словарный лемматизатор, имеющийся в грамматическом словаре, позволяет с намного более высокой точностью находить базовую форму слова из любой грамматической. Если сравнивать результат лемматизации искомого ключевого слова и слов, читаемых из текста, то получается поиск текста с учетом морфологии.

Операция лемматизации позволяет увеличить полноту и точность информационного поиска, т.к. во время операции стемминга отсутствует морфологическая обработка, из-за чего часто в выборку попадают документы, не релевантные запросу, но содержащие совпадающие формы, в то время как в релевантных документах данные слова употребляются в другой форме.

Только с помощью лемматизации можно проводить эффективное индексирование. В этом случае под эффективностью подразумевают скорость индексирования, которая зависит от количества анализируемых слов и их форм. Процесс индексирования идет тем медленнее, чем больше слов приходится обрабатывать.

Благодаря этому возникает идея модифицировать модуль морфологического анализа, в основе которого лежит механизм стемминга. Вместо механизма стемминга использовать механизм лемматизации для быстрого и релевантного поиска ключевых слов.

В Lucene.Net модуль морфологического анализа RussianAnalyzer предназначен

для преобразования слов. Данный модуль используется для разбора, разбивки и индексирования данных, имеет встроенный стоп-лист, а также производит сравнение без учета регистра символов.

Стоп-слова — это слова, не несущие какой-либо самостоятельной смысловой нагрузки. В целях уменьшения базы данных системы не учитывают стоп-слова при индексировании, заменяя их специальным маркером. К ним относятся:

- союзы и союзные слова,
- местоимения,
- предлоги,
- частицы,
- междометия,
- указательные слова,
- цифры,
- знаки препинания,
- вводные слова,
- ряд некоторых существительных, глаголов, наречий (например, однако и др.)

Библиографический список

1. Нистратова, М.В. Алгоритмы поиска, используемые в Lucene.Net / М.В. Нистратова // Вестник МГУЛ – Лесной вестник. – № 6. – 2016. – Т. 19.
2. Нистратова, М.В. Алгоритмы поиска релевантной информации в полнотекстовых базах данных / М.В. Нистратова // Естественные и технические науки. – 2015. – № 10.
3. Нистратова, М.В., Кулагин В.Г. Оценка эффективности поиска документальной информации в системах единой авторизации / М.В. Нистратова, В.Г. Кулагин // Двойные технологии. – № 1. – 2016.
4. Полнотекстовый поиск в веб-проектах: Sphinx, Apache Lucene, Xapian. – Режим доступа: <http://habrahabr.ru/post/30594/>
5. Стемминг и лемматизация. – Режим доступа <http://delaem-krasivo.ru/programmirovanie/234-stemming-i-lemmatizaciya.html>
6. Стемминг. – Режим доступа: <http://gruzdoff.ru/wiki/%D0%A1%D1%82%D0%B5%D0%BC%D0%BC%D0%B8%D0%BD%D0%B3>
7. Стратегии поиска и выдачи информации. – Режим доступа: <http://studall.org/all-130662.html>
8. Тихонов, В. Поисковые системы в сети Интернет / В. Тихонов. – Режим доступа: <http://www.citforum.ru/internet/search/searchsystems.shtml>
9. Шарапов, Р.В. Модели информационного поиска / Р.В. Шарапов, Е.В. Шарапова, Е.А. Саратовцева. – Режим доступа: <http://vuz.exponenta.ru/PDF/FOTO/kaz/Articles/sharapov1.pdf>.
10. Язык запросов Lucene.NET. – Режим доступа: <https://pavelbelousov.wordpress.com/2011/03/23/язык-запросов-lucene-net/>

USE OF THE MODULE OF THE MORPHOLOGICAL ANALYSIS IN LUCENE.NET

Zherdeva M.V., pg. «University of Technology»⁽¹⁾; Artyushenko V.M., Prof. «University of Technology», Dr Sci. (Tech.)⁽¹⁾

masha8908@rambler.ru,

⁽¹⁾«University of Technology», 141070 Moscow region, Korolev, ul. Gagarin, 42

In this article the use of the morphological analysis in Lucene.Net is considered. The relevance of this article doesn't raise any doubts since the developer realizing any project faces a need to carry out a search in the web application. In this article the author considers concepts of a stemming and a lemmatization. The purpose of a stemming and a lemmatization is selected. The use of the module of the morphological analysis is shown in Lucene.Net. Lucene.Net is the retrieval engine postponed from a platform Java in lucene. In lucene – it is high-performance, scalable library for full-text query search. Full-text query search for documents belongs to process of search, information in documents or meta data on documents. In lucene it is allowed to add a search capability in different applications. The principal feature of library is that the small memory size, the existence of the ranged search, a possibility of simultaneous search and up-dating of an index are required. An introduction now and throughout already several years has been the most popular free library of full-text query search. The purposes which need to be reached in case of implementation of the operations executed by the module of the morphological analysis are considered. The idea of modification of algorithm of full-text search Lucene.Net for quick and relevant keywords search engine has been suggested.

Keywords: a full text search, stemming, lemmatization, lemma, stopwords, a token.

References

1. Zherdeva M.V. *Algoritmy poiska, ispol'zuemye v Lucene.Net* [Search algorithms used in lucene.net]. Moscow State Forest University Bulletin – Lesnoy Vestnik. № 6. 2016. T. 19.
2. Nistratova M.V. *Algoritmy poiska relevantnoy informacii v polnotekstovyyh bazah dannyh* [Search algorithm relevant information in the full-text databases]. Estestvennye i tekhnicheskie nauki. 2015, № 10) [Natural and technical science], 2015, No. 10).
3. Nistratova M.V., Kulagin V.G. *Ocenka ehffektivnosti poiska dokumental'noy informacii v sistemah edinoy avtorizacii* [Оценка эффективности поиска документальной информации в системах единой авторизации]. Dvoynye tekhnologii [Двойные технологии]. № 1. 2016.
4. *Polnotekstovyy poisk v veb-proyektakh: Sfinks, Apache Lucene, Xapian* [Full-text search in web projects: Sphinx, Apache Lucene, Xapian.] Available at: <http://habrahabr.ru/post/30594/>
5. *Stemming i lemmatizaciya* [Stemming and lemmatization]. Available at: <http://delaem-krasivo.ru/programmirovanie/234-stemming-i-lemmatizaciya.html>
6. *Stemming* [Stemming]. Available at: <http://gruzdoff.ru/wiki/%D0%A1%D1%82%D0%B5%D0%BC%D0%BC%D0%B8%D0%BD%D0%B3>
7. *Strategii poiska i vydachi informacii* [Strategy of search and issue of information]. Available at: <http://studall.org/all-130662.html>
8. Tihonov V. *Poiskovyie sistemy v seti Internet* [Search engines on the Internet]. Available at: <http://www.citforum.ru/internet/search/searchsystems.shtml>
9. Sharapov R.V., Sharapova E.V., Saratovceva E.A. *Modeli informacionnogo poiska* [Models of information search]. Available at: <http://vuz.exponenta.ru/PDF/FOTO/kaz/Articles/sharapov1.pdf>
10. *Yazyk zaprosov Lucene.NET* [Language of inquiries Lucene.NET]. Available at: <https://pavelbelousov.wordpress.com/2011/03/23/yazyk-zaprosov-lucene-net/>