УДК 303.732.4
ББК 65.05
К 38
А.Ф. Кизянов

Рассмотрение возможности расширения индексации текста на основе стимминга с возможностью автоматической индексации текста на произвольном языке

(Рецензирована)

Аннотация:

В статье рассматривается возможность автоматизации процесса индексирования и поиска текста на естественном языке с применением самообучающегося поискового индекса. Предложена структура электронной библиотеки, удобная для организации поиска. Описан автоматический метод полнотекстового индексирования текста на естественном языке. Показано, что, при наличии определённых свойств языка, предложенный метод индексирования может быть эффективен. Для предложенного метода индексирования описан алгоритм сопоставления проиндексированного текста запросу.

Ключевые слова:

Электронная библиотека, поисковая система, АИПС, релевантность, полнотекстовый поиск, естественный язык, поисковый индекс, анализ морфологии.

Вступление. Электронные библиотеки

В настоящее время, в связи с растущими темпами информатизации общества, всё острее стоит вопрос о переводе в электронный вид огромных массивов информации, хранящейся на бумаге. Развитие так называемых «безбумажных технологий» в областях документооборота и средствах массовой информации, а так же широко распространённое стремление к переносу фонда накопленных человечеством знаний в мир Интернет выдвигают проблему создания эффективных средств перевода “бумажной” информации в «электронную». При возникновении задачи такого перевода, скажем, для библиотеки, фонд которой состоит из сотен тысяч книг, безусловно, необходимо не только произвести сканирование и по необходимости распознание этого объёма текста, но и обеспечить эффективные средства поиска и выбора информации из полученного банка данных на уровне современных технологических возможностей, предоставляемых развитием вычислительной техники. Одной из важнейших областей применения таких технологий является построение электронных научных библиотек.

В общем случае можно сказать, что конечной целью информационного поиска как процесса является отыскание, скорее всего, еще не известных субъекту поиска сообщений (документов), содержащих сведения, нужные для решения стоящих перед ним научных или практических задач. При этом характер информации и способ её представления может быть самым разным – от описания конкретного устройства, предназначенного для непосредственного применения, до некоторой совокупности идей или фактов, приводящих к творческому озарению.

С другой стороны, практика информационного поиска – это рутинный перебор массива документов, сосредоточенных в традиционных или электронных хранилищах (более или менее полно представляющих интересующую нас тему и более или менее структурированных). Отбор обыкновенно проводится по содержанию документов – по значениям реквизитов или поисковым терминам.

Здесь следует уточнить, что на этапе собственно поиска (а не использования найденной информации) под «содержанием» понимается поверхностное представление об изложенной в документе информации – он именно для того и

«отбирается», чтобы изучить его содержание и определить возможность использования воспринятой информации. Кроме того, в реальных системах поиск всегда опосредован: отбор ведется по вторичным документам – поисковым образам, библиографическим и реферативным описаниям. При этом эффективность поиска (по крайней мере сокращение времени необходимого для просмотра и восприятия), обеспечивается за счет систематизации массива по предметному, алфавитному или какому-либо другому принципу.

В этом смысле автоматизированная информационно-поисковая система (АИПС) это комплекс программных и лингвистических средств, обеспечивающих избирательный отбор по заданным признакам документов, хранимых на машиночитаемых носителях обычно в виде баз данных.

При построении поисковых систем для информационных хранилищ такого рода, обозначается ряд особенностей, которые необходимо учитывать разработчикам системы.

–Объём работы не позволяет даже предполагать возможность проведения такой работы вручную, в приемлемое время. Оптимальным выходом, безусловно, остаётся автоматизированное или, лучше, автоматическое выявление удобных для эффективной работы поисковых процессов признаков и построение необходимых для организации поиска индексов.

–Несомненно, такое хранилище должно в качестве кодировки для хранения индексной информации использовать Unicode или другим не менее эффективным способом обеспечивать поддержки многоязычных текстов и во избежание потери ценной, для кого-то возможно бесценной, информации связанной с несовпадением кодировок.

–При организации научной библиотеки немаловажным является так же возможность полноценной работы с текстами на разных языках, а так же – текстами, содержащими термины и цитаты на языках отличных от того, на котором написан сам текст.

–Для электронных массивов информации такого рода (когда терминология для однозначного понимания не должна подвергаться многократным переводам) нежелательно выделение некоторого «родного» языка библиотеки. Текст на всех этапах должен приводиться в том виде, в каком он был написан в оригинале.

Организация индекса библиотеки

Поисковый индекс состоит из языково-тематических блоков, каждый из которых относится к определённому тематическому разделу и индексирует слова на определённом языке.

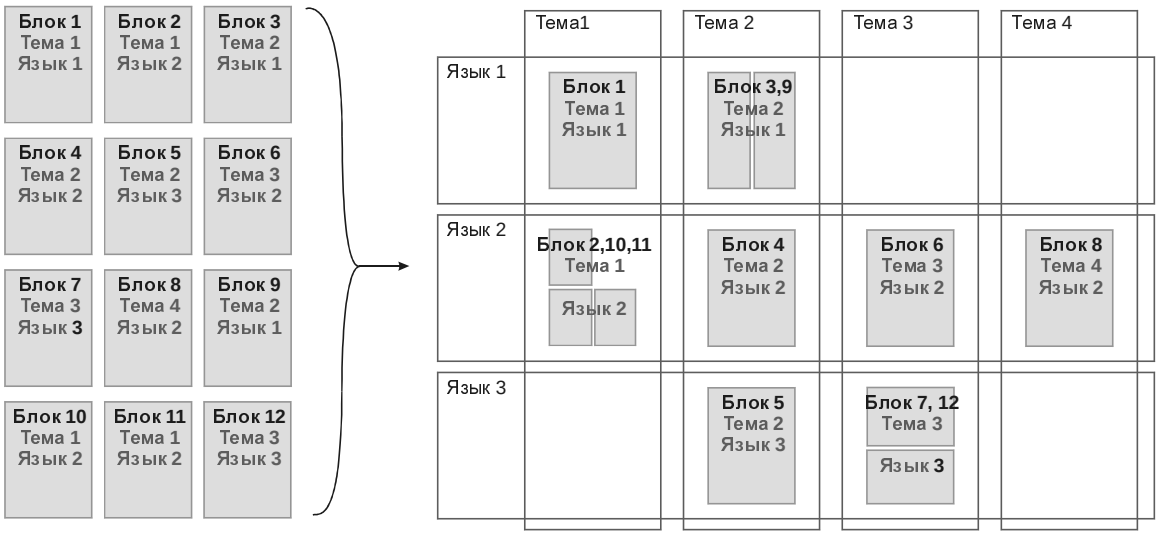

Рисунок 7. Блочная организация поискового индекса библиотеки (пример)

Разделение по блокам происходит по следующим соображениям:

1. Разделение по языку – для разных языков релевантными считаются непересекающие-

ся множества слов (кросс-язычный поиск, в данном случае, не рассматривается).

2. Разделение по теме – разным тематическим разделам соответствуют непересекающиеся множества текстов. Если текст можно отнести к нескольким разделам одновременно, он должен быть отнесён к самому общему из этих разделов по иерархии (в случае, когда используется иерархия разделов).

3. Разделение по размеру – для удобства хеширования размер индекса блока ограничивается 116507 записями. Если в тематико-языковой раздел входит большее количество слов, под раздел выделяются дополнительные блоки.

Каждый блок индекса содержит следующие элементы:

–тезаурус (список слов, индексированных в данном блоке);

–файл инвертированного списка слов (файл, определяющий релевантность слов тезауруса страницам текстов индексированных в данном блоке);

–индекс основных частей слов (каждой основной части слова ставится в соответствие список слов, содержащих данную морфему);

–индексы префиксов и постфиксов (служат для хранения значений коэффициентов релевантности соответствующих морфем).

Каждая проиндексированная книга относится к конкретному тематико-языковому разделу, поэтому все слова, индексированные для конкретной книги записываются в тезаурусы блоков этого языково-тематического раздела.

Структура поискового индекса библиотеки

Поисковый индекс библиотеки организован в виде набора блоков. Каждый блок соответствует определённому языку и тематическому разделу. Размер блока индекса ограничен 116507-ю записями. Ограничение делается с целью упрощения работы с блоком индекса в памяти (один блок занимает 4 Мб оперативной памяти). Каждый блок содержит следующие файлы:

1. Файл инвертированного списка термов (в качестве термов выступают ненормализованные слова, найденные в индексируемых текстах не менее 2-х раз). Этот файл содержит термы блока и для каждого терма – ссылку на соответствующий диапазон в дополнительном файле инвертированного списка термов. Этот диапазон описывает подмножество страниц книг раздела библиотеки, содержащих данный терм ($\mathbf{TP}^i$). Одновременно, файл инвертированного списка термов выступает в роли тезауруса.

2. Дополнительный файл инвертированного списка термов. В этом файле перечисляются ссылки на страницы, на которых встретились термы, ссылающиеся на этот диапазон файла.

3. Файл индекса основных частей слов. Файл содержит список основных частей слов, найденных в процессе кластеризации слов индексируемых текстов. Для каждой основной части слова приводится список слов, содержащих эту морфему ($\mathbf{T}_{mjr}^{k}$). Кроме того, каждой морфеме в этом файле ставится в соответствие числовое значение, определяющее базовый коэффициент релевантности терма, содержащего эту морфему (R_{mjr}^{k}), содержащего эту морфему.

4. Файлы индексов префиксов и постфиксов содержат списки префиксов и постфиксов, соответственно, найденных в процессе кластеризации слов индексируемых текстов. Каждой морфеме в этих файлах ставится в соответствие числовое значение, определяющее вес морфемы при вычислении релевантности слова, содержащего эту морфему (R_{pref}^{l}, R_{postr}^{m}).

Формирование индекса библиотеки

Процесс формирования индекса библиотеки состоит в добавлении в существующий индекс новых книг. При необходимости создаются новые блоки индекса, добавляются новые языки, вводятся новые тематические разделы. Добавление новой книги к индексу делается в 3 этапа

1. Индексация книги – формируется временный постраничный индекс слов книги. Книге даётся уникальный идентификатор. Для индексации многотомных изданий и подшивок и сборников в виде единого источника, используется код раздела. Каждая страница книги при этом определяется так:

`Код_источника(32b).Код_раздела(16b).`
`Номер_страницы(16b)`

2. Дополнение основного индекса библиотеки. При этом

a. Определяется тематический раздел книги.
b. Для каждого слова определённого языка:

i. ищется соответствующий тематико-языковой блок индекса, в который можно ещё добавить слова.
ii. если такой блок не найден, создаётся новый блок.
iii. слово добавляется в тезаурус блока, а в инвертированный список страниц блока вносится информация о том, на каких страницах книги присутствует данное слово.

3. Дополнение морфологической базы блока индекса.

a. дополнение слоя префиксов;
b. дополнение слоя постфиксов;
c. выделение основных частей слов и
d. дополнение слоя основных частей слов.

Заметим, что при индексации книги все слова книги всегда добавляются к блокам одного тематического раздела. Если книга содержит слова на разных языках, для этих языков в этом разделе должны существовать, или быть созданы соответствующие блоки. При этом информация по одной книге может быть распределена по произвольному количеству блоков раздела, но в каждом языковом блоке тематического раздела могут индексироваться только слова этого языка. Таким образом, при поиске данных по книге необходимо использовать все блоки её тематического раздела, но для известного языка запроса можно ограничиться только блоками раздела на этом языке.

Применение индекса на основе расширенного представления термов

При поиске среди слов, представленных в различных склонениях, главное – это правильно определить неизменяемую часть слова (stem), выделив или удалив вспомогательные морфемы. Для этого, в поисковых системах, как правило, применяются специализированные программные средства – стимеры.

Стимер (stemmer) – это программа или алгоритм, позволяющий на основе заранее заложенных в него данных и правил, с помощью знаний (разработчика) о том, как правильно проводить анализ морфологии для слов данного языка, выделить из отдельно взятого слова на этом языке его основную, неизменяемую часть (http://snowball.tartarus.org/).

В качестве широко известного примера стимера можно привести стимер Портера (porter stemmer), предназначенный для нормализации слов английского языка (особенность морфологии английского языка такова, что дополнительные части слов мало информативны и могут быть безболезненно удалены, для русского языка с его развитой морфологией, это менее удобно).

Различают два типа ошибок стимера: перестимминг (owersteming) – удаление вместе со вспомогательными морфемами части символов основы слова, и недо-стимминг (understeming) – включение в основу слова лишних символов.

Для решения этой проблемы, предлагается при индексации сохранять не основу слова, а слово целиком, лишь указав, какую основу это слово содержит. При этом в процессе поиска можно ориентироваться не только на факт наличия в тексте основной части некоторого слова, но и на то, какие вспомогательные морфемы присутствуют в этом слове.

Преимущества такого подхода к индексации в следующем:

–Можно организовать поиск фразы или слова в определённой морфологической форме. При этом факт совпадения вспомогательных морфем для слов фразы будет служить дополнительным признаком похожести текста документа на текст запроса.

–Можно, задавая различные значения веса для вспомогательных морфем, учесть важность тех или иных частей слов (или их комбинаций) в определении релевантности слов друг другу.

–Поскольку при индексации слова сохраняются целиком, есть возможность частично компенсировать последствия пере-стимминга, задав достаточно большое значение коэффициента релевантности для вспомогательных морфем, содержащих части основ слов.

–Варьируя степень влияния коэффициентов релевантности вспомогательных морфем на значение числовой оценки похожести, можно менять поведение поисковой системы от поиска всех слов, содержащих основные части слов запроса, до поиска только слов только в точно той форме, в какой они содержатся в запросе.

Недостатками такого подхода можно считать то, что:

–Происходит увеличение размера индекса засчёт необходимости сохранять в нём дополнительную информацию.

–Усложняется процедура поиска.

Увеличение размера индекса связано с необходимостью хранить помимо списка ссылок на точки индексации надо будет хранить:

а) тезаурус (словарь уникальных слов);

б) списки вспомогательных морфем;

в) информацию о том, какая основная часть слова содержится в каждом отдельно взятом слове тезауруса.

Однако, как и при любом другом индексировании, основное место в индексе займут ссылки на точки индексации (по одной на каждое слово в каждом проиндексированном тексте). Информация о соответствии слов тезауруса основным частям слов имеет числовой характер и может быть представлена компактно. Список основных частей слов должен содержать полный перечень самостоятельных понятий языка, лишённых дополнительных морфологических конструкций (по сути – корни слов) Списки дополнительных морфем вообще не займут сколько-нибудь значительного места. Их в русском языке, например, несколько сотен.

Усложнение процедуры поиска может быть оправдано ростом качества, ради которого и предложен рассматриваемый метод.

Морфологический анализатор блока

Морфологический анализатор блока предлагается реализовать в виде нейронной сети, каждому нейрону которого будет сопоставлена одна морфема – префикс, постфикс или основная часть слова. Нейрон будет, получив на входе некоторое слово, возвращать бинарный признак возможности наличия в этом слове данной морфемы.

Обучение морфологического анализатора блока

При обучении морфологического анализатора предполагается, что все слова в тезаурусе – это слова одного языка. Алфавит языка известен.

Цель обучения – выделить морфемы слов, содержащихся в тезаурусе блока. Невозможно сразу выделить основные части слов, поэтому легче начать с выделения вспомогательных морфем, которые неоднократно встречаются в этих словах и могут быть выявлены статистическим исследованием слов тезауруса. В последствии, основные части могут быть выделены, как части слов, остающиеся после удаления вспомогательных морфем.

При этом можно придерживаться следующей стратегии:

1. Сформировать пустые нейронные слои для префиксов, постфиксов и основных частей.
2. Заполнить слой префиксов морфемами, состоящими из одной буквы алфавита, по одному на каждую букву алфавита.
3. Последовательно подавать на слой префиксов все слова тезаурусов.
4. Удалить все нейроны с нулевым весом.
5. Для нейронов, вес которых превысил некоторое пороговое значение добавить в сеть нейроны-потомки, т.е. нейроны, морфемы которых получаются добавлением к морфеме нейрона-предка одной буквы из алфавита, по одному нейрону-потомку для каждой буквы алфавита (для префиксов буква добавляется справа, а для постфиксов – слева).
6. Повторять шаги 3–5 пока в сеть не стабилизируется.
7. Повторить шаги 2–6 для слоя постфиксов.
8. Заполнить слой основных частей слов основными частями слов тезауруса. Для их выделения для каждое слово подаётся на слои префиксов и постфиксов и за основную часть слова принимается та морфема, которая остаётся после удаления из слова префикса и постфикса нейронов-победителей.

Это обучение ориентируется на получение вариантов ошибочной разбивки слов, представленных в Таблице 2 (префиксы и постфиксы содержат буквы основы слова, основные части не содержат букв вспомогательных морфем). На следующем этапе предполагается провести дообучение сети, чтобы откорректировать разбивки слов на морфемы с целью добиться того, чтобы основные части соответствовали реальным основам слов. При этом следует ограничиться объединением существующих нейронов с соответствующей модификацией их морфем. Таким образом, дообучение должно быть на-

правлено на сокращение количества полученных основных частей слов.

Организация поиска с помощью поискового индекса библиотеки

Для каждого запроса с целью уменьшения вычислительной нагрузки на поисковую систему желательно определить множество тематических разделов, в которых следует производить поиск. Поскольку индексация тематически-языковых разделов производится независимо, целесообразно поиск производить также по каждому разделу по отдельности. Рассмотрим пример поискового алгоритма, осуществляющего поиск страниц (S^{ij}) книг ($\mathbf{S^i}$), входящих в раздел **S**, текст на которых содержит слова (Wq^k), входящие в запрос **Wq**.

Запрос – это набор слов, часть из которых может содержаться в инвертированном списке термов, часть содержит морфемы, содержащиеся в индексе основных частей слов. Остальные слова не будут участвовать в поиске.

Алгоритм поиска может быть таким:

1. Выделение слов запроса из текста запроса **Wq**. На этом этапе из запроса выделяются слова, которые и составят множество термов (Wq^k), участвующих в определении степени релевантности страницы запросу. Множество термов может состоять из слов разных языков. Язык слова определяет то, в каких языковых блоках тематических разделов индекса будет вестись его поиск. При выделении слов запроса желательно отсеять слова, которые заведомо не индексировались при составлении индекса. Для упрощения, дальнейшие рассуждения будут вестись для одного конкретного языка.

2. Построение списков комбинаций морфем $Wq^k <=> \mathbf{W^k} = <Wq_{pref}, Wq_{mjr}, Wq_{post}>^n$ для термов Wq^k запроса, которые будут участвовать в поиске. При этом делается оценка степени соответствия термов Wq^k термам, содержащимся в индексе. Можно выделить три степени соответствия:

a. Поиск термов запроса в тезаурусе. В роли тезауруса выступает предварительно загруженный в оперативную память файл инвертированного списка термов. В случае, если терм присутствует в тезаурусе, ему приписываются те основная часть, префикс и постфикс, которые определены для него в индексе основных частей слов. В поиске для данного терма участвует только эта комбинации морфем (множество $\mathbf{W^k}$ будет состоять из 1-го элемента).

b. Если терм не найден в тезаурусе, производится перебор и проверка на вхождение в данный терм морфем из индекса основных частей слов. Для каждого случая присутствия морфемы в терме производится выделение префикса и постфикса. В поиске для данного терма участвуют все определённые комбинации морфем (множество $\mathbf{W^k}$ будет состоять из n элементов, $n \geq 1$).

c. Если терм не содержит основных частей приведённых в индексе, он в поиске не участвует (множество $\mathbf{W^k}$ будет пустым).

3. Формируется подмножества термов $\mathbf{T^k}$, релевантных слову запроса Wq^k. Эти множества образуются как объединение $\mathbf{T_{mjr}^{kn}}$, соответствующих ему основным частям слов Wq_{mjr}^{kn}, входящим в $\mathbf{W^k}$.

$$T^k = \bigcup_n T_{mjr}^{kn};$$

4. Каждому терму из подмножества $\mathbf{T^k}$ соответствует коэффициент релевантности R^{ki} равный базовому, соответствующему этой основной части коэффициенту релевантности R_{mjr}^{kn}.

$$R^{ki} = R_{mjr}^{kn}$$

5. Если префикс T_{pref}^{ki} (постфикс T_{post}^{ki}) выделяющийся из терма T^{ki} при делении его основной частью $T_{mjr}^{ki} = Wq_{mjr}^{kn}$ совпадает с префиксом W_{pref}^{kn} (постфиксом W_{post}^{kn}) элемента W^{kn} множества **W**, коэффициент релевантности R^{kn} терма T^{ki} вычисляется по формулам:

$$R^{ki} = R_{mjr}^{kn} \times R_{pref}^{kn},$$

$$R^{ki} = R_{mjr}^{kn} \times R_{post}^{kn}$$

или

$$R^{ki} = R_{mjr}^{kn} \times R_{pref}^{kn} \times R_{post}^{kn}.$$

6. Формируются подмножества страниц, релевантных словам запроса $\mathbf{WqP^k}$. Эти множества получаются объединением множеств $\mathbf{TP^{kn}}$, релевантных термам T^k:

$$WqP^k = \bigcup_n TP^{kn}$$

Это множество будет содержать все страницы, на которых есть какие-нибудь слова, «похожие» на слова запроса (похожесть здесь

определяется тем, содержит ли слово ту же основную часть слова, какая есть в одном из слов запроса, а совпадение вспомогательных морфем повышает эту степень похожести).

7. RWP^k – коэффициент релевантности слова Wq^k на странице P, для слов запроса, для которых есть релевантные слова на странице:

$$RWP^k = \underset{i}{MAX}\left(R^{ki}\right);\ P \in WqP^k .$$

Такой подход позволит выбрать самое похожее написание слова на странице, и заведомо отсеет или сделает малозначимыми случаи вхождения в текст на странице мало похожих слов.

8. RP^m – коэффициент релевантности страницы P^m равен сумме RWP^k для слов запроса, для которых есть релевантные слова на странице:

$$RP^m = \sum_j \left(R^j\right);\ P \in WqP^k .$$

В этом случае релевантность страницы будет тем больше, чем большее количество слов запроса с большей степенью похожести было найдено на странице.

9. RBook – коэффициент релевантности книги равен максимальному значению RP^m для страниц, принадлежащих этой книге.

$$RBook^b = \underset{p}{MAX}\left(RP^p\right);\ P \in Book^p .$$

При таком подходе релевантность книги не будет зависит от количества вхождений в книгу релевантных слов, а только от максимально похожей страницы.

Заключение

Выше рассмотрена возможность реализации индексации текста на основе стимминга с возможностью автоматической индексации текста на произвольном языке и предложены алгоритмы построения подобного индекса и поиска по нему.

Предложенный метод индексации позволяет в полуавтоматическом или автоматическом режиме индексировать большие массивы текстовой информации при задании априори минимума информации – алфавита языка, при условии, что индексируемый текст представлен в заранее определённом формате (например Unicode UTF-16 текстовом файле). Принципы построения индекса позволяют индексировать тексты на разных языках. Количество проиндексированной информации не зависит от стандартного размера блока индекса. Разделение блока индекса на часть, загружаемую в оперативную память и часть, постоянно хранимую на вторичном носителе информации и организация взаимодействия со второй частью позволяют минимизировать затраты времени, связанные с обращением ко вторичной памяти.

Примечания:

1. Вишняков Ю.М. Самообучающийся морфологический анализатор для статического корпуса текстов / Кизянов А.Ф. // Сборник трудов VI Всероссийской научной конференции с международным участием «Новые информационные технологии. Разработка и аспекты применения». Таганрог: Таганрогский радиотехнический университет, 2003. сс. 479-482.
2. Вишняков Ю.М. Корректировка разбивки слов на морфемы с помощью самообучающейся нейронной сети / Кизянов А.Ф. // Материалы Международной научно-методической Интернет-конференции «Информационные технологии в образовательной среде современного вуза». Белгород: Изд-во БГТУ им. В.Г. Шухова. 2004. сс. 31-36.
3. Кизянов А.Ф. Организация индекса поисковой системы библиотеки // Материалы II Всероссийской научной конференции молодых учёных, аспирантов и студентов «Информационные технологии, системный анализ и управление». Таганрог. 2004. сс. 89-90.
4. Кизянов А.Ф. Рассмотрение возможности расширения индексации текста на основе стемминга с возможностью автоматической индексации текста на произвольном языке // Перспективные информационные технологии и интеллектуальные системы. №3(19)/2004. http://pitis.tsure.ru/files19/12.pdf