УДК 681.3

ПРОБЛЕМАТИКА И ОРГАНИЗАЦИЯ НЕЧЕТКОГО ПОИСКА

© 2016 М. А. Демихов

Воронежский институт высоких технологий

Работа связана с обсуждением проблематики и организации нечеткого поиска. Указаны возможности использования словарей. Показаны характеристики векторных и кластерных моделей.

Ключевые слова: нечеткий поиск, векторная модель, кластерная модель, словарь.

При организации поиска именных групп (ИГ) в базах персональных данных возникают характерные проблемы, связанные с наличием в запросах орфографических и фонетических ошибок, ошибок ввода информации, а также отсутствием единых стандартов транскрипции с иностранных языков. Вследствие указанных причин задача поиска в базах данных, содержащих ИГ, не может быть в полной мере решена только методами проверки на точное соответствие, в связи с чем становится актуальной задача разработки специальных методов и технологий нечеткого поиска.

В настоящее время методы нечеткого поиска применяются для решения целого ряда задач, таких, как текстовый и мультимедийный поиск, поиск гомологических цепочек макромолекул в молекулярной биологии, поиск в графах и т. д. Однако универсальной методики их решения не существует, поскольку каждая проблема имеет собственную оригинальную специфику. В данной работе предлагается специализированный метод нечеткого поиска фамильно-именных групп, предназначенный для применения в автоматизированных информационных системах, содержащих персональные данные.

На практике часто встает вопрос о поиске среди иностранных имен, уже переданных на русский язык, и, как следствие, о корректной передаче имен собственных с одного языка на другой. При этом одним из основных требований является адекватная передача звучания имени собственного. Известно несколько методов передачи имен. Наиболее распространенным является транслитерация, при которой символу или набору символов из алфавита ставится в соответствие один или несколько символов другого алфавита, причем соответствие проводится скорее по графическому сходству символов. Подобный метод записи позволяет восстановить исходное написание слова, однако не позволяет воспроизвести его звучание лицу, не знакомому с исходным языком. Более удобным является словарный метод, в котором соответствие слов входного языка словам выходного языка задается при помощи некоторого фиксированного словаря. Однако количество имен собственных растет с каждым годом, в связи с чем построить полный и всеобъемлющий словарь на практике не представляется возможным.

Возможен вариант постоянного наращивания объемов словаря, однако он требует привлечения специалистов, знакомых с языком, на постоянной основе. Наиболее удобным является метод транскрипции, в котором звучание слова в одном языке записывается средствами другого (в том числе специализированного) языка. Например, при фонетической транскрипции в качестве выходного может использоваться язык записи фонетики, позволяющий отразить все нюансы произношения слова. Однако подобный язык знаком лишь узкому кругу специалистов, в связи с чем, как правило, используют метод практической транскрипции, в котором звучание слова записывается алфавитом некоторого существующего языка.

Под нечетким поиском понимается поиск по ключевым словам с учётом возможных произвольных ошибок в написании ключевого слова или, напротив, ошибок написания слова в целевом запросе. Основными аспектами организации текстового поиска являются способ построения поискового индекса, выбор поисковой метрики и собственно алгоритмы нечеткого поиска. Структуры данных информационного поиска, как правило, относятся к одной из двух основных групп: векторные и кластерные модели, либо модели на основе ключевых слов.

Основная идея векторных методов состоит в том, что считаются заданными в поисковых слов, и каждый поисковый объект отображается в вектор, называемый профилем, причем величина к-го элемента профи-

Демихов Михаил Александрович – ВИВТ АНОО ВО, магистр, e-mail:msqwfrdemikhovamm@yandex.ru.

ля зависит от частоты вхождения в документ к-го поискового слова, например слова, строки или подстроки длиной n символов (n-граммы). Поисковое выражение также рассматривается, как документ с соответствующим профилем, и ключевым моментом организации поиска является выбор функции корреляции профилей. В выборку попадают документы, для которых корреляционное значение превышает пороговое. Недостатком векторных методов является необходимость считывания профилей всех документов. Устранить этот недостаток можно, разбив все документы на группы (кластеры) и определив в каждом кластере характерного представителя (центроид кластера), что позволяет сначала сравнивать поисковый запрос лишь с центроидами кластеров.

В случае релевантности центроида запросу поиск продолжается далее внутри кластера, причем процесс разбиения выборки на кластеры может быть иерархическим (многоуровневым). При поиске по ключевым словам, как правило, производят выборку всех документов, содержащих хотя бы одно ключевое слово, а затем ранжируют результаты поиска по степени соответствия (релевантности). В основе поиска по ключевым словам лежит использование специализированных индексных словарей двух основных типов. Инвертированный файл (ИФ) – множество пар <ключевое слово, адрес вхождения ключевого слова в документ>. Сигнатурные файлы (СФ) содержат сигнатуры данных, представляющие собой их упрощенные профили, в которых каждый элемент кодируется одним битом. Сжатые ИФ существенно превосходят СФ по производительности для коротких запросов, но проигрывают им на длинных и очень длинных запросах.

Ключевым элементом организации нечёткого поиска является выбор меры сходства слов или обратной функции – функции расстояния между словами, часто называемой метрикой даже в тех случаях, когда она не является метрикой в строгом математическом смысле. Наибольшее распространение здесь получили трансформационные метрики, в области текстового поиска называемые также расстояниями редактирования. Наиболее известны поисковые метрики следующих типов. Расстояние Хемминга между словами одинаковой длины определяется как число позиций, в которых символы не совпадают. Расстояние Левенштейна, позволяющее сравнивать слова различной длины, равно минимальному числу операций преобразования одного слова в другое, причем, допустимы только операции вставки, удаления и замены, которым также присвоена единичная стоимость.

При определении расстояния Дамерау-Левенштейна перестановка символов принимается за единую операцию редактирования с весом 1. Известны меры сходства Джаро и Джаро-Уинклера, представляющие собой нормированные коэффициенты, специально разработанные для сравнения коротких строк, например, компонентов ИГ. Также получили распространение метрики, основанные на подсчёте количества общих подстрок равной длины (n-грамм). Большинство практических алгоритмов поиска ключевого слова в словаре основаны на модификациях одного из следующих известных методов: последовательный поиск (полный перебор всех слов словаря), метод расширения выборки (query extension), метод n-грамм, поиск с использованием хеширования, методы на основе неравенстве треугольника (триангуляционные деревья).

Последовательный поиск предполагает последовательный перебор слов из словаря и сравнение каждого из них с запросом в соответствии с принятой метрикой. Данный метод применяется, например, в системах Agrep и Glimpse. Метод расширения выборки предполагает построение множества всевозможных «ошибочных» слов, например, получающихся из исходного в результате одной операции редактирования Левенштейна, после чего построенные поисковые запросы ищутся в словаре на точное совпадение. Метод широко используется в программах проверки орфографии, например Ispell, и часто связан с применением морфологического анализа, в частности, стемминга.

Метод n-грамм основан на том, что для поиска слов строится инвертированный файл, в котором роль документов играют слова, а роль слов – подстроки длины n, называемые n-граммами. Поиск с использованием хеширования состоит в подборе отображения (хеш-функции) слова, например, во множество чисел или строк, сохраняющего основные характеристики исходного слова и устойчивого к наиболее распространённым ошибкам. Известным примером является хеш-функция Soundex, встроенная в коммерческие СУБД Sybase, MS SQL Server, Oracle. Триангуляционные деревья позволяют индексировать множества произвольной структуры, при условии, что на них задана метрика (не обязательно евклидова). В основу построения триангуляционных деревьев положена идея расположения близких в

смысле заданной метрики объектов в одинаковых поддеревьях.

При поиске в текстовых массивах данный метод менее эффективен, чем при поиске в базах изображений или больших документов. Перейдем теперь к рассмотрению конкретной задачи нечеткого поиска фамильно-именных групп и описанию разработанного для ее решения специализированного метода.

ЛИТЕРАТУРА

1. Карахтанов Д. С. Использование алгоритмов нечеткого поиска при решении задач обработки массивов данных в интересах кредитных организаций / Д. С. Карахтанов // Аудит и финансовый анализ. – 2010. – № 2. URL: www.auditfin.com/2010/2/toc.asp.

2. Бойцов Л. М. Классификация и экспериментальное исследование современных алгоритмов нечеткого словарного поиска / Л. М. Бойцов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды VI Всеросс. научн. конф.(RCDL'2004). – Пущино, Россия, 2004. URL: http://rcdl.ru/doc/2004/paper27.pdf.

3. Карахтанов Д. С. Использование алгоритмов нечеткого поиска при решении задачи устранения дубликатов в массивах данных/ Д. С. Карахтанов // Молодой ученый. – 2010. – Т. 1. – № 8 (19). – С. 150-155.

4. Потапов Е. Н. Нечеткие множества в хранилище данных / Е. Н. Потапов // 2011. URL: http://разработка.рф/blog/? p=346.

5. Рыжов А. П. Модели поиска информации в нечеткой среде / А. П. Рыжов // М.: Изд-во ЦПИ при ММФ МГУ, 2004. – 96 с.

6. Преображенский Ю. П. Некоторые аспекты информатизации образовательных учреждений и развития медиакомпетентности преподавателей и руководителей / Ю. П. Преображенский, Н. С. Преображенская, И. Я. Львович // Вестник Воронежского государственного технического университета. – 2013. – Т. 9. – № 5-2. – С. 134-136.

7. Преображенский Ю. П. Разработка лингвистических средств интеллектуальной поддержки принятия медицинских решений в клинической практике на основе имитационно-семантического моделирования / Ю. П. Преображенский, Н. С. Преображенская, В. В. Ермолова // Information Technology Applications. – 2013. – № 4. – С. 96-114.

8. Преображенский Ю. П. Сравнительный анализ алгоритмов поиска текстовых фрагментов / Ю. П. Преображенский, А. С. Ермаченко // Вестник Воронежского института высоких технологий. – 2010. – № 7. – С. 76-78.

9. Фомина Ю. А. Принципы индексации информации в поисковых системах / Ю. А. Фомина, Ю. П. Преображенский // Вестник Воронежского института высоких технологий. – 2010. – № 7. – С. 98-100.

10. Паневин Р. Ю. Реализация транслятора имитационно-семантического моделирования / Р. Ю. Паневин, Ю. П. Преображенский // Вестник Воронежкого института высоких технологий. – 2009. – № 5. – С. 57-60.

11. Зазулин А. В. Особенности построения семантических моделей предметной области/ А. В. Зазулин, Ю. П. Преображенский // Вестник Воронежского института высоких технологий. – 2008. – № 3. – С. 26-28.

12. Иванов М. С. Разработка алгоритма отсечения деревьев / М. С. Иванов, Ю. П. Преображенский // Вестник Воронежского института высоких технологий. – 2008. – № 3. – С. 31-32.

13. Паневин Р. Ю. Структурные и функцинальные требования к программному комплексу представления знаний / Р. Ю. Паневин, Ю. П. Преображенский // Вестник Воронежского института высоких технологий. – 2008. – № 3. – С. 061-064.

14. Максимова А. А. Анализ методов обработки медицинских данных / А. А. Максимова // Моделирование, оптимизация и информационные технологии. – 2016. – № 2. – С. 5.

15. Мэн Ц. Анализ методов классификации информации в интернете при решении задач информационного поиска / Ц. Мэн // Моделирование, оптимизация и информационные технологии. – 2016. – № 2. – С. 19.

16. Чопоров О. Н. Оптимизация управления функционированием медицинских систем различного уровня / О. Н. Чопоров, И. Я. Львович, К. А. Разинкин, А. А. Рындин // Системы управления и информационные технологии. – 2013. – Т. 53. – № 3. – С. 100-104.

THE PERSPECTIVE AND ORGANIZATION OF FUZZY SEARCH

© 2016 M. A. Demihov

Voronezh institute of high technologies

The paper involves discussion of the problems and organization of fuzzy search. The possibility of using dictionaries is indicated. The characteristics of vector and cluster models are shown.

Keywords: training module, medical facility, software product.