БИБЛИОСФЕРА, 2013, № 2, с. 74–81

Информатика

УДК 025.3:004.424.4
ББК 78.362.5+78.653

ПОСТРОЕНИЕ МОДЕЛИ ПОИСКА В ЭЛЕКТРОННОМ КАТАЛОГЕ БИБЛИОТЕКИ НА ОСНОВЕ НЕЧЕТКОГО ОТНОШЕНИЯ СХОДСТВА

© Л. П. Вершинина[*], М. И. Вершинин[**], А. Ц. Масевич[*], 2013

[*] *Санкт-Петербургский государственный университет культуры и искусств*
191186, г. Санкт-Петербург, Дворцовая набережная, 2/4

[**] *Национальный минерально-сырьевой университет «Горный»*
199106, г. Санкт-Петербург, Васильевский остров, 21 линия, 2

Описывается возможная модель организации поиска в каталогах современных библиотек. Модель построена на основе анализа результатов экспериментального поиска, в котором в качестве запроса используются варианты представления имени «Чайковский» в языках с латинским алфавитом – английском, французском и немецком. Показано, что в данных каталогах имеется инструмент, учитывающий варианты представления имени. Подобный инструмент может быть реализован с помощью предлагаемой модели. Также затрагиваются некоторые вопросы представления имен в разных системах письма, что имеет определенное значение в теории и практике информационного поиска.

Ключевые слова: информационный поиск, математическое моделирование, нечеткие множества, каталоги библиотек, системы письма, стандарты транслитерации.

The paper considers possible model of the search in catalogs of modern libraries. The model is based on the analysis of the results of experimental search, which used as search terms the variants of the name Chajkovskij as it is represented in the languages with Latin alphabet - English, French and German. It is shown that the catalogs of these libraries have an instrument, which allows getting results independent of variants of names representation. This instrument could be built according to the model suggested. In addition, we have touched some questions of representation of personal names in various languages, which is of certain importance in the theory and practice of information search.

Key words: information retrieval, mathematical modeling, fuzzy sets, library catalogs, writing systems, standards of transliteration.

Среди требований, предъявляемых к программному обеспечению современных библиотечно-информационных систем, наличие инструмента учета вариантов написания поискового термина и элемента библиографического описания, являющегося точкой доступа при поиске в каталогах и/или ключевых слов при поиске в полнотекстовых информационно-поисковых системах (ИПС). Такой инструмент может быть обеспечен разными моделями поиска.

В статье описана модель поиска на основе нечеткого отношения сходства. Модель построена на основе анализа результатов экспериментального поиска в каталогах двух национальных библиотек, а также в метапоисковой библиотечной системе – портале «Европейская библиотека» (TEL).

С использованием теории нечетких множеств построена в общем виде математическая модель, на основе которой может быть реализован такой инструмент. Не исключено, что в этих библиотеках или в какой-либо одной из них реализована модель, сходная с построенной нами.

Для исследования использованы каталоги Немецкой национальной библиотеки и Национальной библиотеки Франции; портал Европейской библиотеки, а также авторитетный файл библиотеки Конгресса США. Все указанные системы свободно доступны в сети Интернет.

Системы авторитетного контроля библиотек

При ведении каталогов библиотек всегда возникает проблема унификации элементов каталожной записи. Традиционно эта проблема решается с помощью авторитетного контроля, т. е. базы данных (БД), записи которой учитывают варианты форматов представления и написания различных элементов библиографического описания – личных имен, названий организаций, географических названий, лексики информационно-поисковых языков. Каждая запись определяет принятый вариант соответствующего элемента, а также содержит ссылочные варианты (от которых даются отсылки «см.» к принятому варианту) и варианты, каким-либо образом связанные с объектом (ссылки «см. также»).

Крупнейшими библиотеками мира накоплен многолетний опыт ведения авторитетного контроля. Хорошо известны: система Библиотеки Конгресса США (Library of Congress Authorities), авторитетный файл Немецкой национальной библиотеки (DNB), Gemeinsame Normdatei (GND), авторитетные файлы национальной библиотеки Франции (BNF) Autorités BnF и Répertoire d'autorité-matière encyclopédique et alphabétique unifié (RAMEAU).

Транслитерация кирилловского письма латинским алфавитом

Транслитерация русских текстов латинскими буквами имеет долгую историю и разнообразные традиции как в России, так и в других странах, где используются различные языки и различные системы письма [6]. В настоящее время действует большое количество международных и национальных стандартов (табл. 1).

Содержание табл. 1 не исчерпывает всех существующих стандартов транслитерации. В ней приведены в основном стандарты, использующиеся в издательской и библиотечной практике. Кроме них существует, например, стандарт транслитерации для заграничных паспортов российских граждан, стандарт для международных телеграмм, водительских прав и др.

Следует помнить, что для представления иностранных имен используется не только транслитерация, но и транскрипция, т. е. фиксация на письме звучания слова.

В реальности эти два подхода (транслитерация и транскрипция) сосуществуют, а порой смешиваются. Так, передача на русском языке имен Ивлин Во (Evelyn Waugh) или Шарль Бодлер (Charles Baudelaire) является примером транскрипции. С другой стороны, при передаче на русском языке имени Генрих Гейне (Heinrich Heine) отчасти использовалась транслитерация, по-немецки, как известно, это имя звучит «Хайнрих Хайне». Однако классик немецкой литературы русским читателям известен именно как Гейне. Кроме того, в этом случае имя «Генрих» имеет традиционное написание по-русски. По-видимому, латинское «н» прежде транслитерировалось русской буквой «г». Таким образом, в данном случае имеет место смешение традиционного написания и устаревшей транслитерации.

Английская фамилия Huxley, которая сегодня устойчиво транскрибируется как «Хаксли» (известный писатель Олдос Хаксли), в конце XIX века транслитерировалась как Гукслей (биолог Thomas Huxley – родственник писателя), а несколько позже – Гексли, вариант представления имени, который, строго говоря, не является ни транслитерацией, ни транскрипцией, а просто произвольной передачей.

Выбор варианта транслитерации (или транскрипции) зависит от многих факторов, в первую очередь, от фонетических систем целевых языков. Необходимо также отметить, что транслитерация имен имеет диахронический аспект, т. е. меняется во времени. Эти изменения хорошо прослеживаются на графиках, построенных с помощью системы Ngram Viewer Google (http://books.google.com/ngrams). Система Ngram Viewer Google позволяет определять частоту встречаемости формы слова в нескольких миллионах книжных текстов на разных языках и строить графики изменения этой величины во времени.

Таблица 1

Действующие системы и стандарты транслитерации русского алфавита латиницей

Обозначение	Описание
ISO 9A и **ISO 9B**	Международный стандарт ISO 9-95 (ГОСТ 7.79–2000). Ранние версии: ISO/R 9:1954, ISO/R 9:1968, ISO 9:1986, взамен ГОСТ 16876–71
ALA-LC (ALA–LC)	Система транслитерации Американской ассоциации библиотек и Библиотеки Конгресса ALA–LC. Используется библиотеками США, Канады и Великобритании (1976 г., обновл. 1997 г.)
BSI	Британский институт стандартов (British Standards Institution & Chemical Abstracts Service BS 2979 (1958 г.)
BGN	Стандарт, принятый комиссией по географическим названиям США (BGN, 1944 г.) и постоянным комитетом по географическим названиям Великобритании (PCGN, 1947 г.)
UN	Стандарт Организации Объединенных Наций (ООН) для географических названий (1987 г.)
WT	Международная научная система транслитерации (Wissenschaftliche Transliteration)
DIN (DIN 1460)	Немецкий институт по стандартизации совместно с Немецкой библиотекой (1982 г.)
Duden	Немецкий стандарт практической транскрипции от Dudenverlag (1991 г.)

Для нашего исследования мы отобрали лишь один тип вариантов написания – разночтения, которые возникают при транслитерации личных имен с кирилловского алфавита на латиницу (выбрано имя «Чайковский Пётр Ильич»). Замечено, что русские слова, в которых встречаются согласные «ч», «ж», «ш», «щ» и двойные гласные «е», «ё», «ю», «я» представляют трудность при транслитерации на латиницу.

Для построения графиков выбраны пять наиболее распространенных вариантов транслитерации (1. Tchaikovsky, 2. Chaikovsky, 3. Cajkovskij, 4. Čajkovskij, 5. Tchaikowsky).

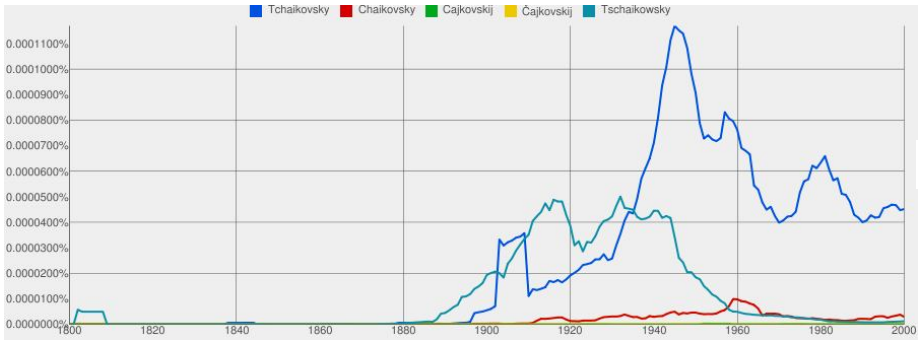
В настоящее время в России принят стандарт ГОСТ 7.79–2000 (см. табл. 1), который является русской версией международного стандарта ISO 9-95.

ГОСТ 7.79–2000 предлагает два варианта транслитерации – строгую (система А), при которой один знак исходного языка заменяется только одним знаком целевого языка (ГОСТ 7.79–2000, таблица 2 для неславянских языков), и ослабленную (система Б), при которой знак исходного языка может быть представлен более чем одним знаком целевого языка [4].

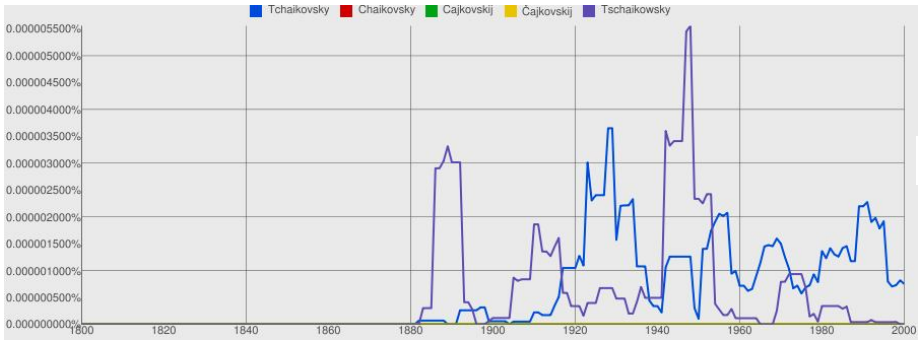
Таким образом, по этим стандартам фамилия «Чайковский» в транслитерации латиницей может быть представлена в двух вариантах: Čajkovskij (система А), Chajkovskij (система Б).

На представленных ниже графиках (рис. 1) отчетливо видны, по крайней мере, четыре закономерности:

1. В каждом из трех языков присутствуют одновременно более одного варианта транслитерации.
2. В разных языках преобладают разные варианты.
3. Частота встречаемости каждого варианта меняется со временем. Отмечается появление новых вариантов.
4. В английском и французском языках в разные временные периоды преобладали варианты 1 (Tchaikovsky) и 5 (Tchaikowsky), а в немецком языке отмечается устойчивое преобладание варианта 5.



а) английский язык



б) французский язык

в) немецкий язык

Рис. 1. Распространенность различных вариантов написания фамилии Чайковский (1800–2000 гг.)

Отметим, что вариант 4, транслитерированный по системе А и, следовательно, соответствующий международному стандарту ISO 9-95, ни в одном из трех естественных языков не встречается (рис. 1–3). Это вполне естественно – знак «Č» в этих языках не используется. Тем не менее, как будет показано ниже, именно эта форма принята национальными библиотеками Германии и Франции в качестве основной.

Отражение личного имени «П. И. Чайковский» в авторитетных файлах трех национальных библиотек

Авторитетные записи на личное имя «П. И. Чайковский» найдены в авторитетном файле библиотеки Конгресса США, национальной библиотеки Франции и Немецкой национальной библиотеки.

Принятой формой транслитерации в авторитетном файле библиотеки Конгресса США (Library of Congress Authorities) выбрана форма Tchaikovsky, Peter Ilich, 1840–1893 (http://lccn.loc.gov/n79072979). Эта форма транслитерации, как видим, не соответствует стандарту ISO 9-95. Запись дает всего 51 вариант написания имени, в том числе вариант в русском алфавите, вариант на иврите, а также варианты в арабской и китайской системах письма. Таким образом, запись содержит 47 вариантов транслитерации латиницей.

В авторитетном файле Немецкой национальной библиотеки (GND) в качестве принятой указана форма Čajkovskij, Pëtr I. (http://d-nb.info/gnd/118638157), форма транслитерации соответствует стандарту ISO 9-95, в записи имеется 85 вариантов написания только латиницей.

В авторитетном файле национальной библиотеки Франции Autorités BnF в качестве принятой выбрана форма Čajkovskij, Petr Il´ič (1840–1893) (http://catalogue.bnf.fr/ark:/12148/cb13900329p/PUBLIC). Транслитерация соответствует ISO 9-95 за исключением буквы «ё», которая в соответствии со стандартом транслитерируется как «ё». В записи содержится 9 вариантов написания латиницей.

Результаты экспериментального поиска в каталогах Немецкой национальной библиотеки и национальной библиотеки Франции

Из авторитетных файлов трех библиотек отобрано 10 вариантов имени, причем 3 варианта полного представления и 7 вариантов только фамилии, после чего по отобранным вариантам произведен поиск в каталогах библиотек и сводных каталогах, доступных по протоколу Z39.50 через портал «Европейская библиотека» (TEL) (http://theeuropealibrary.org). Результат поиска приведен в табл. 2.

Из табл. 2 видно, что положительный результат поиска получен в среднем в 16 библиотеках, причем во многих библиотеках найдено несколько форм записей.

Из библиотек, в которых получен положительный результат, отобраны две – DNB и BNF. В каждой из них положительный результат получен при поиске по 8 вариантам.

В каталогах этих двух библиотек был проведен поиск по нескольким вариантам написания имени.

Таблица 2

Результат поиска в портале «Европейская библиотека»

Форма представления имени	Источник (авторитетный файл библиотеки)	Общее число найденных записей	Число библиотек, в которых получен положительный результат*
Tchaikovsky, Peter Ilich	БК США	10 450	19
Tchaîkovsky, Piotr Ilitch	БК США	5	2
Tchaikovsky, Pyotr Ilyich	БК США	10 262	17
Cajkovskij	BNF	10 799	23
Tchaïkovsky	BNF	20 129	23
Čajkovskij	BNF	11 125	17
Tchaikovsky	BNF	16 496	24
Tschaikowsky	DNB	12 941	20
Chaikovsky	DNB	7 693	13
Chaikovski	DNB	99	4

* Среднее число библиотек, где получен положительный результат: 16,2.
Под положительным результатом понимается результат поиска, в котором имеется хотя бы одна запись.
Примечание. БК – Библиотека Конгресса.

Мы исходили из предположения, что если результаты поиска в каталоге будут близки по числу найденных записей, то, следовательно, такой каталог имеет программный инструмент, возможно основанный на логике нечетких множеств, который позволяет осуществлять извлечение записи независимо от формы написания элемента записи, который является точкой доступа при поиске (табл. 3, 4).

Таблица 3

Результаты поиска по различным вариантам имени в каталоге DNB

Форма написания имени	Число найденных записей
Tchaikovsky, Peter Ilich	6 960
Tchaikovsky, Pyotr Ilyich	7 122
Cajkovskij	7 214
Tchaïkovsky	8 503
Čajkovskij	7 214
Tchaikovsky	8 503
Tschaikowsky	10 290
Chaikovsky	6 991
Среднее число записей на один вариант	7 849,63
Максимальное число (Tschaikowsky)	10 290
Минимальное число (Tchaikovsky, Peter Ilich)	6 960

Таблица 4

Результаты поиска по различным вариантам имени в каталоге BNF

Форма написания имени	Число найденных записей
Tchaikovsky, Piotr Illitch.	5 360
Tchaïkowsky, Piotr Illitch	5 365
Czajkowski, Piotr	5 361
Tchaîkovsky, Piotr Ilitch	5 386
Čajkovskij, Petr Il´ič	5 383
Tschaikowsky	424
Среднее число записей на один вариант	4 546,5
Максимальное число (Tchaîkovsky, Piotr Ilitch)	5 386
Минимальное число (Tschaikowsky)	424

Из табл. 3 видно, что массивы записей на каждый вариант количественно близки. Различия в числе найденных записей при поиске только по фамилии могут быть обусловлены наличием в каталоге однофамильцев композитора.

Из табл. 4 видно, что результаты поиска в каталоге BNF отличаются от результатов в DNB. Разброс при поиске по различным формам полного имени, в разных форматах, разброс между числом записей не превышает 25. Поскольку в пяти из шести запросов использован полный формат имени, то, по-видимому, отсутствует фактор омонимии, т. е. однофамильцев.

При поиске по варианту транслитерации Tschaikowsky извлечено значительно меньшее число записей, что, вероятно, связано с тем, что эта форма последние 20 лет мало употребляется во французском языке (см. рис. 1, *б*), в противоположность немецкому языку (см. рис. 1, *в*), где с 1920 г. эта форма остается наиболее употребительной.

Данные нашего экспериментального поиска подтверждают, таким образом, наличие в каталогах этих двух библиотек инструмента, обеспечивающего относительную независимость результатов поиска от варианта написания.

Отметим, что наш эксперимент носит предварительный (пилотный) характер и данные нуждаются в проверке и статистической обработке.

Некоторые теоретические разработки поиска в библиотечных каталогах с использованием теории нечетких множеств

Национальные библиотеки характеризуются большим объемом фондов, многоязычием. При создании, усовершенствовании и эксплуатации поискового аппарата электронных каталогов таких библиотек разработчики, библиотекари и пользователи имеют дело с задачами и процессами, в которых необходимо оперировать нечеткими понятиями и знаниями. К таким задачам можно отнести автоматизацию индексирования документов, поиск с учетом многоязычия и различий в написании слов и т. п.

Для построения соответствующих алгоритмов (классификации, поиска записей) нечеткие понятия и знания необходимо формализовать. Эту задачу – описание нечетких понятий и знаний, оперирование многозначными и/или неполно определенными лексическими единицами и в конечном итоге построение моделей поиска, соответствующих потребностям пользователей, позволяет решить теория нечетких множеств.

Преимущество нечеткого подхода состоит также в том, что в рамках многозначной логики осмысленные решения находятся для более широкого класса проблем, нежели при четкой постановке.

Так М. И. Вершининым [1] предлагается использовать тезаурус с нечеткими связями между

его элементами (нечеткий тезаурус) для уменьшения затрат при ведении и повышении эффективности использования электронного каталога.

В работе М. И. Вершинина, Л. П. Вершининой [2] предлагается алгоритм автоматической классификации (индексирования) библиографических записей при вводе их в электронный каталог. В основе алгоритма лежит идея автоматической нечеткой классификации (индексирования) записей. Каждая библиографическая запись должна быть отнесена к определенному классу, обозначенному определенной предметной рубрикой. Принадлежность записи тому или иному классу (иначе говоря, присвоение записи индекса в виде той или иной предметной рубрики) определяется в ходе автоматической классификации на основе сравнения лексики записи с имеющимися представителями кластеров – ключевых слов, при этом отношение сходства является нечетким. Классы не имеют четко выраженных границ, в связи с чем принадлежность записи тому или иному классу часто не всегда определяется однозначно.

При нечеткой классификации каждый документ может быть отнесен к нескольким классам с разной степенью принадлежности. Несколько документов будут отнесены к определенному классу, если степень принадлежности каждого документа данному классу максимальная по сравнению со степенями принадлежности другим классам.

Таким образом, появляется возможность организовать поиск в массивах информации по нечетким признакам.

Для решения проблемы поиска и коррекции ошибок системы, а также поиска с учетом наличия ошибок разработан метод нечеткого сравнения строк, основанный на использовании аппарата теории нечетких множеств [3]. Методы нечеткой логики позволяют работать в условиях недостатка статистических данных и сравнивать строки с учетом возможного наличия ошибок без коррекции строк и вмешательства оператора. В разработанном методе учитываются как характер возможных ошибок, так и их ранжирование по частоте появления и другим критериям.

Теоретический подход к построению модели поиска с применением теории нечетких множеств

Формализация нечетких понятий и отношений обеспечивается введением лингвистической и нечеткой переменных, нечеткого множества и нечеткого отношения.

В теории нечетких множеств и нечеткой логике нечеткие отношения играют очень важную роль. Традиционно нечеткие отношения находят приложения в задачах моделирования структуры сложных систем, в управлении технологическими процессами, при анализе процессов принятия решений. Что касается проблем ведения и эксплуатации библиотечных электронных каталогов, то для их решения нечеткие отношения практически не используются. Ниже мы покажем перспективность использования нечетких отношений для организации эффективного поиска в электронных каталогах.

Теория нечетких отношений используется при качественном анализе взаимосвязей между объектами исследуемой системы. При этом учитываются различия в силе связей между объектами.

Обычно нечеткое n-арное отношение R определяется как подмножество декартового произведения n множеств [5]:

$$R \subseteq X_1 \times X_2 \times ... \times X_n$$

и задается с помощью функции принадлежности

$$\mu_R : \quad X_1 \times X_2 \times ... \times X_n \to L .$$

В качестве L может быть взято, например, множество вещественных чисел, отрезок вещественной прямой, множество лингвистических переменных, множество m-мерных векторов, псевдобулева алгебра, полная дистрибутивная решетка и т. п.

Такой подход в определении L дает возможность строить различные обобщения понятия отношения, которые могут использоваться в самых разных областях и, кроме того, позволяет в результате интерпретации различных функций со значениями из L как нечетких отношений применять для анализа свойств этих функций хорошо развитый аппарат теории отношений.

Что касается организации поиска в электронных библиотечных каталогах, то использование нечетких отношений позволяет с единой точки зрения рассмотреть множество факторов, влияющих на качество поиска, в частности, определить связь между запросом и записями в каталоге с учетом многих факторов.

Для иллюстрации возможностей решения некоторых проблем поиска в каталогах библиотек ограничимся рассмотрением бинарных нечетких отношений.

В общем случае, нечетким бинарным отношением R между множествами X и Y называется функция

$$R : \quad X \times Y \to L ,$$

где L – полная дистрибутивная решетка, т. е. частично упорядоченное множество, в котором любое непустое подмножество имеет наибольшую нижнюю и наименьшую верхнюю грани, и операции пересечения $\wedge$ и объединения $\vee$ в L удовлетворяют законам дистрибутивности. Все операции над нечеткими отношениями определяются с помощью этих операций из L.

Если в качестве L взять ограниченное множество вещественных чисел, то операциями взятия

наибольшей нижней и наименьшей верхней грани будут, соответственно, операции inf и sup, операциями пересечения $\wedge$ и объединения $\vee$ будут операции min и max. Эти операции будут определять и операции над нечеткими отношениями.

В случае, когда L является отрезком вещественной прямой [0,1], функция R будет записываться в виде функции принадлежности

$$\mu_R : \quad X \times Y \to [0,1].$$

Если множества X и Y конечны, нечеткое отношение R между X и Y можно представить с помощью его матрицы отношения, строкам и столбцам которой ставятся в соответствие элементы множеств X и Y, а на пересечении строки x и столбца y помещается элемент R(x; y).

В случае, когда множества X и Y совпадают, нечеткое отношение R называется нечетким отношением на множестве X. Такому отношению можно поставить в соответствие взвешенный граф, в котором каждая пара вершин (x; y) из X соединяется стрелкой с весом R (x; y).

Модель поиска в библиотечном каталоге с использованием нечеткого отношения сходства

Пусть для поиска в каталоге выбрано имя русского композитора (автора) – «Чайковский Петр Ильич».

Авторитетная запись на данное имя из авторитетного файла некоторой библиотеки представляет собой множество вариантов написания имени в транслитерации латинским алфавитом, которое обозначим X. Тогда возможные элементы множества X выглядят следующим образом:

x_1 – Tchaikovsky,
x_2 – Tchaïkovsky,
x_3 – Čajkovskij,
x_4 – Tschaikowsky,
x_5 – Chaikovsky и т. п.

Пусть количество вариантов равно m, т. е. $\dim X = m$.

Построим нечеткое отношение сходства между записями с учетом различных факторов.

1. Учет вариантов транслитерации (фактор 1).

Нечеткое отношение сходства записей R_1 представим матрицей сходства:

$$M_1 = \{\mu_1(x_i; x_j)\},\ i, j = 1,\ldots, m,$$

где $\mu_1(x_i; x_j)$ – оценка степени сходства записей; $\mu_1(x_i; x_j) \in [0,1]$.

Матрица сходства может быть получена как в результате количественной оценки некоторого параметра, отражающего связь между записями (количество совпадающих знаков, порядок их следования и т. п.), так и в результате опроса экспертов, которые для каждой пары записей из X указывают их степень сходства в некоторой шкале сравнений, состоящей, например, из фраз типа: «очень сильное сходство», «сильное сходство», «сходство средней силы», «слабое сходство» и т. д. Очевидно, что матрица M_1 в общем случае будет несимметрична.

2. Учет распространенности различных вариантов транслитерации имени в различных языках (фактор 2).

Для учета данного фактора используем данные диахронического исследования вариантов написания имени Чайковский (рис. 1).

На основе данных диахронического исследования построим матрицу:

$$A = \{a(x_i; t_j)\},\ i = 1,\ldots, m,\ j = 1,\ldots, n,$$

где $a(x_i; t_j)$ – оценка степени распространенности формы транслитерации имени x_i в году t_j.

Тогда нечеткое отношение сходства записей R_2 с учетом фактора 2 можно представить матрицей сходства:

$$M_2 = A \bullet A^T,$$

где A^T – транспонированная матрица.

Отметим, что матрица M_2 имеет размерность $m \times m$ и является симметричной.

3. Пусть число факторов, которые необходимо учесть при построении нечеткого отношения сходства, равно k. Определив матрицы нечетких отношений M_1, M_2, …, M_k, строим матрицу M нечеткого отношения сходства R, учитывая все факторы:

$$M = M_1 \wedge M_2 \wedge \ldots \wedge M_k.$$

Используем теперь построенное нечеткое отношение для поиска записей в электронном каталоге по запросу z.

Алгоритм поиска выглядит следующим образом:

4. Устанавливается степень сходства запроса z с записями множества X. В результате получаем вектор:

$$\mu_z = \{\mu_z(x_1), \mu_z(x_2), \ldots, \mu_z(x_m)\},$$

где $\mu_z(x_i)$ – оценка степени сходства запроса z и записи x_i (i = 1, 2,…, m).

Для получения оценки степени сходства может быть использован алгоритм нечеткого сравнения строк [3]. Отметим, что алгоритм работает и в случае, когда запрос содержит ошибки.

5. Из множества записей X выбираем запись x_{i0}, такую, что:

$$\mu_я(x_{i0}) = \max_i \mu(x_i), \quad i = 1,2,\ldots,m.$$

6. В матрице M нечеткого отношения R выбираем строку с номером i_0, соответствующую записи x_{i0}. Эта строка дает возможность ранжировать все

записи из X по степени сходства с x_{i0}, т. е. по степени сходства с запросом z. По сути, получаем откорректированный на основе нечеткого отношения записей каталога вектор μ_z. Результаты поиска начинают выдаваться с записи x_{i0} в соответствии с результатами ранжирования.

При поиске можно ввести порог α на силу нечеткого отношения сходства R, например, $\alpha = 0{,}5$, установив, таким образом, выбор наиболее значимых записей.

Поиск начинается с максимальной по степени сходства записи. Заметим, что матрица сходства M в общем случае является несимметричной, поэтому, начиная поиск с разных записей, мы получаем разный результат.

Заключение

Из таблиц 2, 3 видно, что число записей незначительно различается в зависимости от формы имени. Описанный выше алгоритм позволяет не только получить аналогичный результат, но и улучшить характеристики поиска за счет учета большего числа факторов.

В статье описана модель поиска с использованием нечеткого отношения сходства (матриц сходства). Она была инициирована результатами нашего экспериментального исследования. Предлагаемая модель может оказаться эффективной при разработке поискового аппарата электронного каталога.

Кроме этого, в ходе работы получены некоторые дополнительные результаты. Представляется, что наблюдаемая корреляция между результатами поиска в каталогах по различным вариантам транслитерации, и использованием этих форм в естественном языке (в печатных документах) может представлять определенный интерес в дальнейших исследованиях.

Литература

1. *Вершинин М. И.* Создание нечеткого тезауруса для электронного каталога // Информационные ресурсы библиотек и их кадровое обеспечение : материалы Междунар. науч.-практ. конф., 23–26 мая 2000 г. – Минск, 2000. – С. 91–96.
2. *Вершинин М. И., Вершинина Л. П.* Применение нечеткой логики в гуманитарных исследованиях // Библиосфера. – 2007. – № 4. – С. 43–47.
3. *Вершинин М. И.* Электронный каталог: проблемы и решения. – СПб. : Профессия, 2009. – 232 с.
4. ГОСТ 7.79–2000. Правила транслитерации кирилловского письма латинским алфавитом. – М. : Госстандарт, 2001. – 20 с.
5. Нечеткие множества в моделях управления и искусственного интеллекта / под ред. Д. А. Поспелова. – М. : Наука, 1986. – 312 с.
6. *Реформатский А. А.* О стандартизации транслитерации латинскими буквами русских текстов // Науч.-техн. информ. Сер. 1. – 1972. – № 10. – С. 32–36.
7. *Zakharov V. P., Masevich A. C., Pimenov E. N.* Authority control as a linguistic support element of an automated library system // International cataloguing and bibliographic control. – 1996. – Vol. 25, N 4. – P. 84–86.

Материал поступил в редакцию 14.11.2012 г.

Сведения об авторах: *Вершинина Лилия Павловна – доктор технических наук, профессор, заведующий кафедрой информатики и математики, тел.: (812) 496-37-56, e-mail: gukikiim@mail.ru, kiim@spbguki.ru,*

Вершинин Михаил Иосифович – кандидат педагогических наук, доцент кафедры механики, тел.: (812) 328-82-00, e-mail: rectorat@spmi.ru,

Масевич Андрей Цезаревич – старший преподаватель кафедры информационного менеджмента, тел.: (812) 315-16-16, e-mail: kavbibved@mail.ru