Таким образом, задача снова сводится к известным подходам, основанным на модели избыточности.

В связи с этим сравним основные характеристики существующих методов сжатия и метода, основанного на применении псевдорегулярных чисел (ПРЧ), описанного в работе автора [4] (см. таблицу).

Метод сжатия с использованием псевдорегулярных чисел является универсальным и может служить основой для построения систем резервного копирования информации в глобальных информационных системах, а также систем обмена информационными кластерами. При переходе к квантовым технологиям метод позволяет создать хранилища данных (storage) с новой технологией представления информации.

СПИСОК ЛИТЕРАТУРЫ

1. Габидулин Э. М., Пилипчук Н. И. Теоремы Шеннона для источника // Лекции по теории информации. М.: Изд-во МФТИ, 2007. С. 49–52.

2. Методы сжатия данных. Устройство архиваторов, сжатие изображений и видео / Д. А. Ватолин, А. В. Ратушняк, М. В. Смирнов, В. А. Юкин. М.: Диалог-МИФИ, 2002. С. 384.

3. Сэломон Д. Ю. Сжатие данных, изображения и звука. М.: Техносфера, 2004. С. 368.

4. Воробьев Е. Г., Цехановский В. В. Псевдорегулярные числа в двоичных полях // Изв. СПбГЭТУ «ЛЭТИ». 2014. № 2. С. 18–22.

E. G. Vorobiev
Saint-Petersburg state electrotechnical university «LETI»

COMPRESSION OF BINARY CODES ON THE BASIS OF TRADITIONAL METHODS AND USE OF PSEUDO-REGULAR NUMBERS

In article the comparative analysis of the existing methods of data compression and new, on the basis of use of numbers with pseudo-regular binary structure is carried out. The approach allows to solve a problem of storage of reserve information of large volumes that is characteristic for cloudy structures and cluster systems of data-processing centers are offered.

Data compression, pseudo-regular binary structure, methods and algorithms of compression, reduction of impact on volume of the stored information

УДК 681.3, 004.031.4

М. М. Заславский, Т. А. Берленко
ООО «Fruct» (Санкт-Петербург)

Реализация механизма подбора рекомендаций в информационной системе «Открытая Карелия»

Рассмотрен гибкий подход к построению рекомендаций на основании баллов близости для информационной системы с анонимным доступом без использования информации о пользователе и его оценках содержимого этой системы. Даны определения основных понятий подхода. Приведены примеры формул для вычисления баллов близости при сравнении содержимого по данным различной природы (полнотекстовые поля, теги, поля с конечным множеством значений). Описана программная реализация системы подбора рекомендаций для ИС «Открытая Карелия», приведены ее ограничения и направления для ее дальнейшего улучшения.

Рекомендательные системы, баллы близости, системы с анонимным доступом

На сегодняшний день системы подбора рекомендаций стали одной из важнейших частей вебсайтов различной направленности. Примерами могут служить онлайн-каталог кинофильмов Imdb [1], онлайн-магазин Amazon [2], видеохостинг Youtube [3]. Одной из причин широкого

использования подобных систем является возможность удержания пользователя на сайте предоставлением дополнительной информации, адекватной предпочтениям пользователя и содержимому текущей веб-страницы [4]. Наиболее эффективный результат достигается в том случае, если для рекомендательной системы доступны данные о поведении пользователя на сайте (статистика посещенных страниц, время пребывания, оценки и оставленные комментарии). Однако в ряде случаев специфика веб-сайта такова, что он подразумевает только анонимный доступ (без идентификации отдельных пользователей) и поэтому данный подход использовать нельзя. В таком случае одним из возможных решений является вычисление рекомендаций на основе данных самих рекомендуемых страниц. Применение данного подхода к конкретной информационной системе требует решения определенных задач, а именно – выбор критерия близости объектов, обработка неструктурированных данных, проверка корректности рекомендаций. Перечисленные задачи не имеют универсального решения для различных предметных областей, поэтому построение рекомендательной системы для сайта с анонимным доступом является актуальной задачей.

Информационная система «Открытая Карелия». Информационная система «Открытая Карелия» была реализована в 2014 г. в рамках российско-финляндского проекта «Еврорегион Карелия: музейный гипертекст» [5]. Цель системы – предоставление доступа к экспонатам и экспозициям музеев российской и финляндской Карелии. Задачи, решаемые системой:

1. Хранение данных.
2. Ввод данных.
3. Предоставление API для доступа и обработки данных.
4. Предоставление веб-фронтендов для пользователей системы.

Так как в проекте участвует большое количество музеев, предоставляемая информация имеет разный формат [5], поэтому для ее хранения используется NoSql решение MongoDb [6]. Базовой сущностью в рассматриваемой информационной системе являются Объекты. Объект представляет собой набор текстовых полей, описывающих реальный музейный экспонат, здание, персоналию, аудио- или видеоматериал, статью, изобразительный материал или документ. Единственным обязательным полем объекта является поле «Имя», остальные поля могут образовывать достаточно произвольные наборы. Помимо подробной информации об оцифрованных музейных объектах данные системы включают в себя также интерактивные планы экспозиций, тематические наборы объектов (истории), теги, мультимедиа-контент. Все возможные поля объектов ИС «Открытая Карелия» приведены на рис. 1. Для упрощения рассмотрения полем «Дата» обозначена вся совокупность хранимой в БД информации о времени и дате создания, открытия или изготовления реаль-

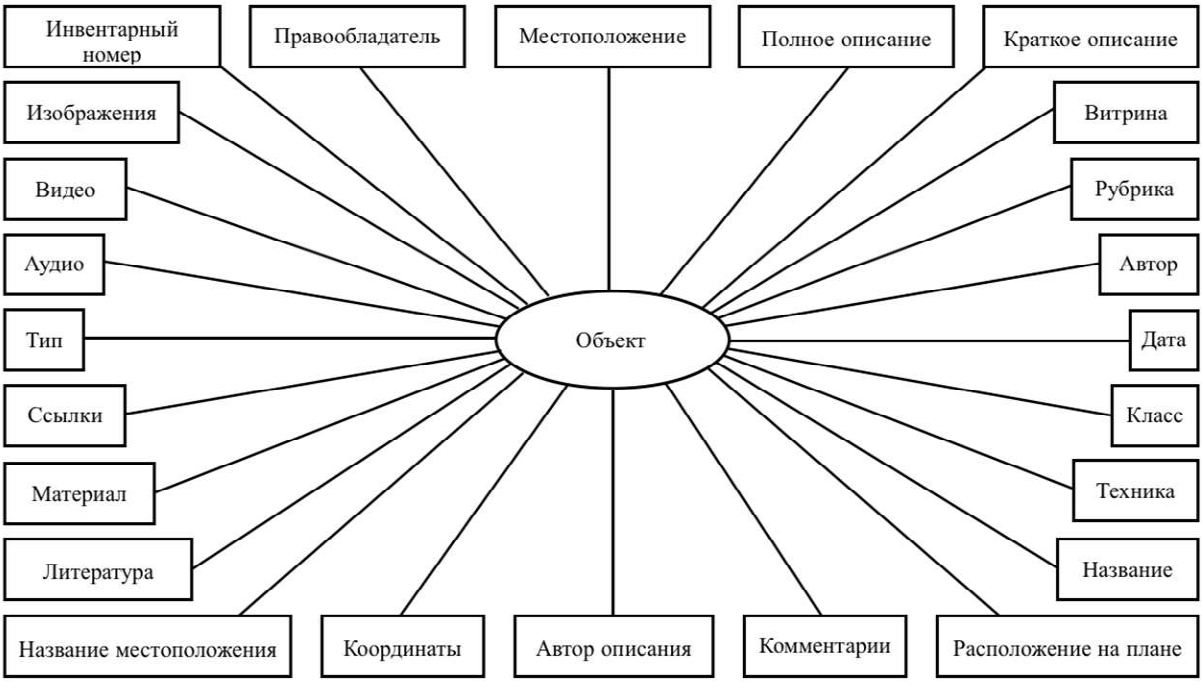

Рис. 1

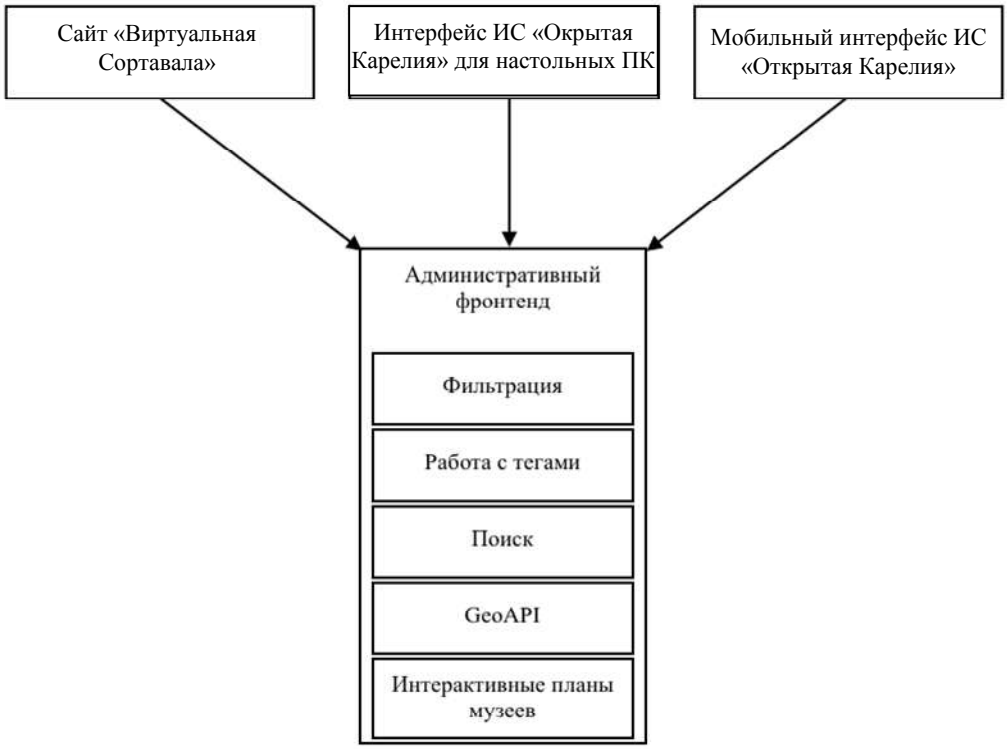

Рис. 2

ного экспоната или объекта, соответствующего объекту системы. Для обработки геоданных «Открытая Карелия» использует открытую LBS-платформу Geo2Tag [7], [8].

Информационная система «Открытая Карелия» предоставляет API для работы с тегами объектов, включающий в себя разметку и фильтрацию объектов по тегам, вывод статистики. Ввод возможных тегов в систему осуществляется вручную, а разметка объектов – автоматически с использованием словарей словоформ.

Интерфейс доступа пользователей к данным «Открытой Карелии» реализуется на основе веб-фронтендов, использующих веб-приложения, написанные на языке Python с помощью библиотек Flask и Jinja2. На данный момент система имеет 3 фронтенда для конечных пользователей. Общая схема ресурсов системы представлена на рис. 2.

Центральным элементом каждого фронтенда является «Карточка объекта» – страница с подробным описанием конкретного объекта. Помимо текстовых полей самого объекта данная страница содержит также информацию о тегах объекта, связанных мультимедиа-файлах и историях, географическом расположении объекта.

Доступ на страницу «Карточка объекта» предоставляется двумя путями – через сгенерированный системой QR-код, который сканирует пользователь, и непосредственный переход на карточку объекта со страниц фронтендов. Кроме того, доступ к веб-фронтендам ИС «Открытая Карелия» не подразумевает регистрации и авторизации.

Постановка задачи. Необходимо разработать универсальный интерфейс для построения списка рекомендаций по заданному объекту ИС «Открытая Карелия», удовлетворяющий следующим требованиям:

1. Интерфейс должен быть гибким с точки зрения замены критерия близости объектов.

2. Вычисление близости объектов должно учитывать индивидуальные множества тегов и полнотекстовые поля, такие, как «Название», «Полное описание», «Аннотация».

Выбор источника данных для построения рекомендаций. Прежде чем решать задачу непосредственно построения рекомендательной системы, необходимо определиться с подходом к подбору рекомендаций. На сегодняшний день в литературе описаны следующие источники данных для построения рекомендаций в ИС [4], [9]:

1. Использование данных об оценках содержимого ИС пользователями.

2. Использование информации о пользователе.

3. Использование информации о содержимом.

4. Использование информации о пользователе и содержимом.

Несомненно, что источники данных 1–3 позволяют достигнуть высокой степени персонифицированности построенных рекомендаций за счет

прямого учета явно (через информацию о пользователе) или неявно (через оценки содержимого) указанных интересов пользователя. Однако их использование в ИС «Открытая Карелия» невозможно, так как система реализует только анонимный доступ к содержимому ИС. В связи с этим для решения поставленной задачи источником был выбран подход, использующий только информацию о содержимом (объектах) системы.

Разработанное решение. Используемый подход. Для построения рекомендаций предлагается использовать подход, основанный на вычислении баллов близости между всеми объектам системы для различных критериев близости. *Баллами близости* будем называть неотрицательное число, характеризующее сходство двух объектов в смысле определенного критерия близости. Значения баллов близости объекта **A** и объекта **B** в смысле критерия близости **C** прямо пропорциональны сходству объекта **B** с объектом **A** в рамках критерия близости **C**. *Критерием близости* будем называть совокупность критерия отбора и упорядоченного и взвешенного набора полей (*список полей*), по которому проводиться сравнение. *Критерий отбора* – это ограничение на поля объекта **B**, в случае невыполнения которого значение баллов близости объекта **A** и объекта **B** в смысле критерия близости **C** равно нулю.

Рассмотрим процедуру вычисления значения баллов близости объектов **A** и **B** в смысле критерия близости **C**. Обозначим значение баллов близости как **D**. Шаги процедуры:

1. Проверка соответствия критерию отбора. Если объект **B** не соответствует критерию отбора, то значение баллов близости берется равным нулю и процедура завершается.

2. Для каждого поля из списка полей критерия близости определяется сходство значений у объектов **A** и **B**. Значение степени сходства, умноженное на значение весового коэффициента поля, прибавляется к текущему значению **D**. Процедура вычисления численной степени сходства различных полей будет описана далее.

3. Если объекты **A** и **B** входят одновременно в одну или более историй, к значению **D** добавляется значение *d*, умноженное на количество общих историй.

Достоинства описанного подхода:

– гибкость – манипулируя критериями отбора, составом и весовыми коэффициентами списков полей критерия можно получить критерий близости, отражающий связь между объектами по различным параметрам;

– универсальность – поля разного типа учитываются единообразно при построении рекомендаций.

Типы рекомендаций. Как было показано ранее, ИС «Открытая Карелия» не может использовать данные о конкретных пользователях и их предпочтениях для построения рекомендаций, поскольку использование системы подразумевает только анонимный доступ к пользовательским фронтендам. Для того чтобы учесть различные интересы различных пользователей, были предложены несколько критериев, рекомендации по которым демонстрируются во фронтенде одновременно. В таблице перечислены названия критериев близости, используемых в системе, и их критерии отбора. Списки полей не приводятся ввиду их неинформативности и большого объема.

Название	Критерий отбора
Объекты из других музеев	Поле «Музей» у **A** и **B** не совпадает
Объекты с похожими тегами	Множества тегов **A** и **B** имеют непустое пересечение
Объекты с похожей датой	Год, вычисленный из поля «Дата» объектов **A** и **B**, совпадает
Объекты с похожим местом	Поле «Музей» у **A** и **B** не совпадает
Похожие объекты из недвижимого наследия	Поле «Класс» объекта **B** равно «Недвижимое наследие»
Объекты с другим классом	Поле «Класс» у **A** и **B** не совпадает
Рекомендации	Поле «Класс» у **A** и **B** совпадает

Вычисление баллов близости для различных полей. Рассмотрим, каким образом вычисляются значения степени сходства для отдельных полей объектов ИС «Открытая Карелия». Будем считать, что в критериях близости используются только следующие поля: название, теги, расстояние, класс, рубрика, местоположение, автор, витрина, автор описания, место, тип, техника, материал, расположение на плане. Приведенные поля были отобраны, так как они предоставляют всесторонние характеристики объекта зрения, используя минимум информации.

Для поля «Название» предлагается вычисление функции n_{dist}, обратно пропорциональной расстоянию Левенштейна [10] между значениями данного поля для сравниваемых объектов:

$$n_{\text{dist}}(n_1, n_2) = \frac{1}{\text{lev}(n_1, n_2) + 1},$$

где n_1 и n_2 – значения поля «Название» для первого и второго сравниваемых объектов; $\text{lev}(n_1, n_2)$ – функция вычисления расстояния Левенштейна между двумя строками. Выбор функции такого вида обусловлен тем, что ее свойства обусловливают понимание сходства названий объектов в смысле баллов близости. Наибольшее значение функции n_{dist} приходится на случай, когда строки названий полностью совпадают и расстояние Левенштейна равно нулю. По мере роста различия между названиями и, как следствие, увеличения расстояния Левенштейна значение n_{dist} асимптотически стремится к нулю, что соответствует полностью различным объектам в смысле баллов близости.

Для поля «Теги» предлагается использовать мощность пересечения множеств тегов сравниваемых объектов:

$$t_{\text{dist}}(t_1, t_2) = |t_1 \cap t_2|.$$

Использование подобного функционала позволяет напрямую количественно учесть степень сходства двух объектов по множествам тегов в смысле баллов близости – в случае отсутствия общих тегов значение функционала равно нулю, а при наличии общих тегов оно будет прямо пропорционально их количеству. Очевидным недостатком данной формализации является унификация различных тегов – так, например, при наличии четко выраженной связи между двумя объектами формализация одинаковым образом учтет как теги, отражающие данную связь, так и теги, к данной связи не относящиеся.

Для поля «Дата» количественная степень сходства вычисляется с использованием количественной разницы значений данного поля у обоих объектов, обозначенной через D:

$$d_{\text{dist}}(d_1, d_2) = \frac{1}{D + 1}.$$

Значение D определяется в зависимости от имеющейся информации о датах для сравниваемых объектов по следующей процедуре:

– если оба объекта имеют информацию о годе (веке) в поле «Дата», то D вычисляется как разница годов (веков);

– если у одного объекта в поле «Дата» отсутствует информация о годе, но есть информация о веке, в то время как у другого есть оба значения, то разница D вычисляется как разница веков для объектов.

Для прочих полей объекта в качестве количественной степени сходства используется точное сравнение значений – в случае равенства значений количественная степень равна 1, иначе 0. Подобный подход был выбран по двум причинам. Во-первых, точное сравнение полей существенно проще в реализации, чем описанные ранее подходы на основе теоретико-множественных операций или расстояния Левенштейна. Во-вторых, множества значений всех прочих полей (кроме полей «Название», «Теги», «Дата») являются конечными и имеют низкие мощности.

Реализация механизма построения рекомендаций. Механизм построения рекомендаций был реализован в виде программного модуля для ИС «Открытая Карелия». Работа с рекомендациями в рамках модуля разделена на 2 этапа:

1. Построение кеша рекомендаций. При каждом запуске бэкэнда «Открытой Карелии» происходит последовательное вычисление значений баллов близости для всех пар объектов системы и всех критериев близости. Вычисленные значения индексируются совокупностью упорядоченной пары идентификаторов пары объектов и идентификатора критерия близости.

2. Предоставление программного интерфейса получения рекомендаций по идентификатору объекта и критерия близости. Подборка рекомендованных объектов осуществляется выбором N объектов, которым соответствуют N наибольших значений баллов близости при заданном критерии близости.

Использование предварительного кеширования обусловлено прежде всего требованиями производительности, так как вычисление баллов близости в процессе обращения пользователя к системе может привести к существенной задержке ее ответа. Однако такой подход сопряжен с существенными ограничениями:

1. Объем требуемой памяти прямо пропорционален количеству объектов в БД.

2. Для обновления данных в кеше необходим перезапуск веб-приложения.

3. При запуске веб-приложения сервис рекомендаций не доступен на время построения кеша.

Разработанная система построения рекомендаций для ИС «Открытая Карелия» представляет собой решение, основанное на понятии баллов близости. Данное решение является гибким – для построения нового типа рекомендаций необходим только критерий отбора, набор полей и весов, по которым будут вычисляться баллы близости. При этом система учитывает поля различной структуры – имеющие фиксированное количество значений, полнотекстовые и теги, что позволяет находить взаимосвязи между различными по своей природе объектами.

Дальнейшая работа над системой рекомендаций будет включать в себя:

– исследование применимости и целесообразности различных критериев близости;

– разработку расширений системы для сравнения полнотекстовых полей большого объема («Краткое описание», «Полное описание»);

– реализацию механизма оффлайн-кеширования – построения кеша рекомендаций фоновой процедурой с периодическим обновлением данных.

СПИСОК ЛИТЕРАТУРЫ

1. Personalized Recommendations Frequently Asked Questions. URL: http://www.imdb.com/help/show_leaf?personalrecommendations.

2. Amazon.com Help: About Recommendations. URL: http://www.amazon.com/gp/help/customer/display.html/ref=hp_left_sib?ie=UTF8&nodeId=16465251.

3. Рекомендованные YouTube. URL: https://www.youtube.com/feed/recommended, свободный.

4. Еврорегион Карелия: Музейный Гипертекст. URL: http://openkarelia.org/about, свободный.

5. Rokach Lior, Bracha Shapira, Paul B. Kantor. Recommender systems handbook. Vol. 1. New York: Springer, 2011.

6. MongoDB. URL: http://www.mongodb.org/, свободный.

7. Romanikhin V., Zaslavsky M. Spacial Filters For Geo2tag LBS Platform // Proc. of 11th Conf. of Open Innovations Association FRUCT, St.-Petersburg, Russia, 23–27 Apr. 2012. P. 151–157.

8. Geo2Tag Implementation for MAEMO / I. Bezyazychnyy, K. Krinkin, M. Zaslavskiy, S. Balandin, Y. Koucheravy // Proc. of 7th Conf. of Open Innovations Framework Program FRUCT, St.-Petersburg, Russia, 26–30 Apr. 2010. P. 7–11.

9. Melville P., Mooney R., Nagarajan R. Content-Boosted Collaborative Filtering for Improved Recommendations // AAAI-02, Austin, TX, USA. 2002. С. 187–192.

10. Левенштейн В. И. Двоичные коды с исправлением выпадений, вставок и замещений символов. // Докл. АН СССР. 1965. Т. 163, вып. 4. С. 845–848.

M. M. Zaslavskiy, T. A. Berlenko
LLC «Fruct» (Saint-Petersburg)

IMPLEMENTATION MECHANISM OF SELECTION RECOMMENDATIONS IN INFORMATION SYSTEM «OPEN KARELIA»

A flexible approach for construction of recommendations based on points of closeness for informational system with anonymous access and without use of information about a user and his assessments of contents of this system is considered. Definitions of basic ideas of the approach are given. Examples of formulas for calculation of closeness points in view of comparison of contents with data of different nature are put (full-text fields, tags, fields with eventual quantity of values). Program realization of selection system of recommendations for IS «Open Karelia» is described, contingencies for it and directions for its further improvement are set.

Recommendatory systems, points of proximity, systems with anonymous access