УДК 004.05

ИСПОЛЬЗОВАНИЕ ТЕОРЕТИКО-МНОЖЕСТВЕННОГО ПОДХОДА ДЛЯ ПОИСКА НЕОБХОДИМОГО КОНТЕНТА ПО АТРИБУТАМ И КЛЮЧЕВЫМ СЛОВАМ

Соколова Е.А.

ФГБОУ ВПО «Северо-Кавказский горно-металлургический институт (Государственный технологический университет)», Владикавказ, e-mail: katya_sea@mail.ru

При разработке аппаратно-программного комплекса музеев и картинных галерей стояла задача определить оптимальный подбор экспонатов. Для реализации поставленной задачи рассматривается теоретико-множественный подход. Теоретико-множественный подход предполагает анализ множеств информационных сообщений, объектов и т.п. с точки зрения их количественных признаков. Содержание поиска качественного аспекта информации заключается в том, чтобы выделять, изучать и исследовать характеристики множества сообщений в связи с качественными моментами составляющих его частей. Одним из путей решения задачи по превращению потенциальной информации в информацию актуальную является использование наиболее рациональных средств кодирования (декодирования) информации. В статье описана формальная постановка для решения поставленной задачи, алгоритм поиска и рассмотрен пример решения задачи поиска необходимого контента по атрибутам и ключевым словам.

Ключевые слова: теоретико-множественный подход, множество информационных сообщений, контент

USE SET-THEORETIC APPROACH TO FIND YOUR CONTENT OVER THE ATTRIBUTES AND KEYWORDS

Sokolova E.A.

North Caucasian Institute of Mining and Metallurgy (State Technological University), NCIMM (STU) Vladikavkaz, e-mail: katya_sea@mail.ru

In the development of the hardware-software complex of museums and art galleries task was to implement the best selection of items. To accomplish the task, consider the set-theoretic approach. The set-theoretic approach involves the analysis of sets of data messages, objects, etc. in terms of quantitative traits. Search the qualitative aspect of the content of the information is to select, to explore and investigate the characteristics of multiple messages in connection with quality moments of its constituent parts. One way to solve the problem of turning the potential of information in the information up to date is to use the most efficient means of encoding (decoding) of information. This article describes a formal statement to the task, the search algorithm and is considered an example of solving the problem of finding the necessary content based on attributes and keywords.

Keywords: set-theoretic approach, a set of data communications content

При разработке аппаратно-программного комплекса музеев и картинных галерей стояла задача реализовать оптимальный подбор экспонатов. Разработанные программные средства должны поддерживать поиск необходимого контента по атрибутам и ключевым словам.

Для решения этой задачи был использован теоретико-множественный подход.

Теоретико-множественный подход

Теоретико-множественный подход предполагает анализ множеств информационных сообщений, объектов и т.п. с точки зрения их количественных признаков. Здесь не происходит полного отмежевания от качества исследуемых информационных объектов и их элементов. Напротив, уже само исследование формализованных множеств, сообщений (например, данных судебной статистики и др.) постоянно предполагает наличие каких-либо качественных моментов, признаков, позволяющих говорить о содержании исследуемых информационных сообщений. Содержание поиска качественного аспекта информации заключается в том, чтобы выделять, изучать и исследовать характеристики множества сообщений в связи с качественными моментами составляющих его частей.

Одним из путей решения задачи по превращению потенциальной информации в информацию актуальную является использование наиболее рациональных средств кодирования (декодирования) информации (например, определение в цифровом коде ЕГРПОУ информации о предприятии или выражение текста закона в условных символах специального информационно-поискового языка).

Формальная постановка задачи

Для обозначения эффективного алгоритма поиска оптимальных экспонатов выставочного центра в ответ на запрос пользователя удобно использовать теоретико-множественный подход. Исследуемые объекты представим в виде множества их

свойств, которые определены для оценки в процессе поиска:

$$O = \{p\} = \{p_0, p_1, \ldots, p_n\} \qquad (1)$$

где O – исследуемый объект; p – свойство, которое участвует в поиске.

Приоритетность свойств при поиске и оценке релевантности отображается в виде множества весовых коэффициентов, которые определяются путем экспертного анализа исследуемого объекта (в данном случае – экспоната выставки).

$$\{k\} = \{k_0, k_1, \ldots, k_n\}. \qquad (2)$$

Поисковый запрос соответственно теоретико-множественному подходу удобно представить как множество слов. Кроме того, для повышения эффективности и точности поиска, а также для обеспечения высокой достоверности результатов из множества слов поискового запроса удаляются все повторения и семантически «слабые» конструкции (например, союзы и предлоги).

$$R = \{w, \forall w(F_{sem}(w) \neq \emptyset)\}, \qquad (3)$$

где R – поисковый запрос; w – слово в запросе; $F_{sem}(w)$ – функция определения семантических соответствий для слова.

Функция определения соответствия свойства исследуемого объекта слову из запроса определяется как произведение соответствующего весового коэффициента и показателя оценочной функции вхождения слова в значение свойства:

$$F_{eval}(p, w) = k_p \cdot F_{ex}(p, w), \qquad (4)$$

где p – свойство объекта; w – слово в запросе; k_p – весовой коэффициент для свойства; F_{ex} – оценочная функция встречаемости слова в значении свойства.

Оценочная функция встречаемости слова зависит от способа интерпретации значения свойства и целей поиска. Например, для атрибута «Автор произведения» сущности «Экспонат» оценочная функция встречаемости слова может быть определена системой следующего вида:

$$F_{ex}(p_{auth}, w) = \begin{cases} 1, & \text{если} \quad w \in val(p_{auth}) \\ 0, & \text{если} \quad w \cap val(p_{auth}) = \varnothing \end{cases}, \qquad (5)$$

где p_{auth} – свойство «Автор произведения»; w – слово поискового запроса; $val(p_{auth})$ – множество значений свойства «Автор произведения» для каждого объекта.

Используя представление (1) и функцию (4), интегральный показатель соответствия исследуемого объекта поисковому запросу (релевантность) определяем как среднее арифметическое суммы показателей функции (4):

$$S(\{p\}, \{w\}) = \sum_{i=1}^{n} \frac{\sum_{j=1}^{m} F_{eval}(p_i, w_j)}{m}. \qquad (6)$$

Результаты оценки релевантности для соответствующих объектов и сами исследуемые объекты представляются в виде множества, элементы которого сортируются по убыванию уровня релевантности.

Алгоритм поиска

1. Даны множества свойств объектов $O = \{p\} = \{p_0, p_1, \ldots, p_n\}$ и множества весовых коэффициентов $\{k\} = \{k_0, k_1, \ldots, k_n\}$, определяющих приоритетность свойств.

2. Вводим в строку поиска запрос R, включающий в себя слова $w_i\ (i = 1, m)$.

3. Определяем соответствие свойства исследуемого объекта слову из запроса по формуле

$$F_{eval}(p, w) = k_p \cdot F_{ex}(p, w),$$

где $F_{eval}(p, w) = k_p \cdot F_{ex}(p, w)$ равно 1, если слово входит в свойство, 0 в противном случае.

4. Определяем среднее арифметическое суммы показателей функции $F_{eval}(p, w) = k_p \cdot F_{ex}(p, w)$

$$S(\{p\}, \{w\}) = \sum_{i=1}^{n} \frac{\sum_{j=1}^{m} F_{eval}(p_i, w_j)}{m}.$$

5. Сортируем объекты в порядке убывания уровня релевантности, определенного на шаге 4.

6. Вывод объектов.

Пример решения задачи поиска необходимого контента по атрибутам и ключевым словам

Дана табл. 1, содержащая некоторое количество экспонатов.

Шаг 1.

Задаем множество свойств для оценки в процессе поиска

$$O = \{p\} = \{p_0, p_1, \ldots, p_n\}, \qquad (1)$$

где O – каждый исследуемый объект; p – свойство, которое участвует в поиске; p_1 – название; p_2 – автор; p_3 – год создания; p_4 – описание; p_5 – категория.

$$O = \{p\} = \{\text{название, автор, год создания, описание, категория}\}.$$

Таблица 1

Экспонаты

№ п/п	Название	Автор	Год создания	Описание	Категория
1	Марфа Посадница	Дмитрий Иванов	1808	Вручение пустынником Феодосием Борецким меча Ратмира юному вождю новгородцев Мирославу, назначенному Марфой Посадницей в мужья своей дочери Ксении	Живопись
2	Церковь в тумане	Леонардо да Винчи	1515	После сотворения портрета «Мона Лиза» к последним годам жизни относится туринский автопортрет Леонардо	Импрессионизм
3	Мона Лиза	Леонардо да Винчи	1515	Портрет госпожи Лизы Джокондо. По-итальянски ma donna	Живопись
4	Единоборство князя Мстислава	Андрей Иванов	1803	Единоборство князя Мстислава Владимировича Удалого с косожским князем Редедей	Живопись

Приоритетность свойств при поиске и оценке релевантности отображаем в виде множества весовых коэффициентов

$$\{k\} = \{k_0, k_1, \ldots, k_n\};$$

$$\{k\} = \{1; 1; 0{,}2; 0{,}2; 0{,}2\}$$

Шаг 2.

Введем исходный поисковый запрос, включающий в себя строку «Мона Лиза».

Шаг 3.

Определим уровень релевантности для первого объекта:

$$F_{eval}(p_1, \text{«Мона»}) = 0;$$
$$F_{eval}(p_2, \text{«Мона»}) = 0;$$
$$F_{eval}(p_3, \text{«Мона»}) = 0;$$
$$F_{eval}(p_4, \text{«Мона»}) = 0;$$
$$F_{eval}(p_5, \text{«Мона»}) = 0;$$
$$F_{eval}(p_1, \text{«Лиза»}) = 0;$$
$$F_{eval}(p_2, \text{«Лиза»}) = 0;$$
$$F_{eval}(p_3, \text{«Лиза»}) = 0;$$
$$F_{eval}(p_4, \text{«Лиза»}) = 0;$$
$$F_{eval}(p_5, \text{«Лиза»}) = 0;$$

$$S_1\left(\{p\}, \{w\}\right) = \sum_{i=1}^{n} \frac{\sum_{j=1}^{m} F_{eval}(p_i, w_j)}{m} = 0.$$

Данный результат не будет отображен в результатах поиска.

Шаг 4.

Определим уровень релевантности для второго объекта:

$$F_{eval}(p_1, \text{«Мона»}) = 0;$$
$$F_{eval}(p_2, \text{«Мона»}) = 0;$$
$$F_{eval}(p_3, \text{«Мона»}) = 0;$$
$$F_{eval}(p_4, \text{«Мона»}) = 0.2;$$
$$F_{eval}(p_5, \text{«Мона»}) = 0;$$
$$F_{eval}(p_1, \text{«Лиза»}) = 0;$$
$$F_{eval}(p_2, \text{«Лиза»}) = 0;$$
$$F_{eval}(p_3, \text{«Лиза»}) = 0;$$
$$F_{eval}(p_4, \text{«Лиза»}) = 0.2;$$
$$F_{eval}(p_5, \text{«Лиза»}) = 0;$$

$$S_1\left(\{p\}, \{w\}\right) = \sum_{i=1}^{n} \frac{\sum_{j=1}^{m} F_{eval}(p_i, w_j)}{m} = 0.2.$$

Данный результат войдет в список отображаемых результатов поиска.

Шаг 5.

Определим уровень релевантности для третьего объекта:

$$F_{eval}(p_1, \text{«Мона»}) = 1;$$
$$F_{eval}(p_2, \text{«Мона»}) = 0;$$
$$F_{eval}(p_3, \text{«Мона»}) = 0;$$
$$F_{eval}(p_4, \text{«Мона»}) = 0;$$
$$F_{eval}(p_5, \text{«Мона»}) = 0;$$
$$F_{eval}(p_1, \text{«Лиза»}) = 1;$$
$$F_{eval}(p_2, \text{«Лиза»}) = 0;$$
$$F_{eval}(p_3, \text{«Лиза»}) = 0;$$
$$F_{eval}(p_4, \text{«Лиза»}) = 0;$$
$$F_{eval}(p_5, \text{«Лиза»}) = 0;$$

$$S_1\left(\{p\}, \{w\}\right) = \sum_{i=1}^{n} \frac{\sum_{j=1}^{m} F_{eval}(p_i, w_j)}{m} = 1.$$

Данный результат в список отображаемых результатов поиска.

Шаг 6.

Определим уровень релевантности для четвертого объекта:

$$F_{eval}(p_1, \text{«Мона»}) = 0;$$
$$F_{eval}(p_2, \text{«Мона»}) = 0;$$
$$F_{eval}(p_3, \text{«Мона»}) = 0;$$
$$F_{eval}(p_4, \text{«Мона»}) = 0;$$
$$F_{eval}(p_5, \text{«Мона»}) = 0;$$
$$F_{eval}(p_1, \text{«Лиза»}) = 0;$$
$$F_{eval}(p_2, \text{«Лиза»}) = 0;$$
$$F_{eval}(p_3, \text{«Лиза»}) = 0;$$
$$F_{eval}(p_4, \text{«Лиза»}) = 0;$$
$$F_{eval}(p_5, \text{«Лиза»}) = 0;$$

$$S_1\left(\{p\},\{w\}\right) = \sum_{i=1}^{n} \frac{\sum_{j=1}^{m} F_{eval}(p_i, w_j)}{m} = 0.$$

Данный экспонат не войдет в список отображаемых результатов поиска.

Шаг 7.

Результат поиска (табл. 2).

Таблица 2

Результаты поиска

№ п/п	Название	Автор	Год создания	Описание	Категория
3	Мона Лиза	Леонардо да Винчи	1515	Портрет госпожи Лизы Джокондо. По-итальянски ma donna	Живопись
2	Церковь в тумане	Леонардо да Винчи	1515	После сотворения портрета «Мона Лиза» к последним годам жизни относится турин-ский автопортрет Леонардо	Импрес-сионизм

Критическое обсуждение результатов

В результате разработки аппаратно-программного комплекса музеев и картинных галерей был реализован оптимальный подбор экспонатов с использованием теоретико-множественного подхода. Разработанные программные средства поддерживают поиск необходимого контента по атрибутам и ключевым словам.

Список литературы

1. Знаменский А.В., Черкалин С.Д. Компьютер в экспозиции. Взгляд из провинции. // Музеи и информационное пространство: проблема информатизации и культурное наследие. Тезисы докладов. Пятая ежегодная конференция АДИТ-2001. – Тула, 2001. – С. 15–16.

2. Соколова Е.А., Гречаный С.В.Оптимизация алгоритма компрессии видеоизображений вариабельными фрагментами // Устойчивое развитие горных территорий. – 2011. – № 4(10). – 5 с.

3. Соколова Е.А., Мирошников А.С. Разработка программных продуктов для конвертации мультимедийных изображений // Перспективы науки. – 2012. – № 11(38).

4. Черненко В.В. Использование автоматизированных информационных систем в экспозиции – с. 88 Музеи Москвы и музеология XX века: Тезисы научной конференции (М., 25–26 ноября 1997 г.) // отв. ред. Ю.У. Гуральник – М.: РГГУ, 1997.

5. Эльзассер М.Э. Князева Н.А. Новое измерение партнерства: виртуальная выставка и реальное сотрудничество // Электронный потенциал музея: стимулы и ограничения, достижения и проблемы: тезисы докладов XXX Международной конференции CIDOC-АДИТ-2003. – СПб., 2003. – С. 55–56.

References

1. Znamenskij A.V., Cherkalin S.D. Komp'juter v jekspozicii. Vzgljad iz provincii. // Muzei i informacionnoe prostranstvo: problema informatizacii i kul'turnoe nasledie. Tezisy dokladov. Pjataja ezhegodnaja konferencija ADIT-2001. Tula, 2001 pp. 15–16.

2. Sokolova E.A., Grechanyj S.V.Optimizacija algoritma kompressii videoizobrazhenij variabel'nymi fragmentami// Mezhdunarodnyj nauchnyj zhurnal-Ustojchivoe razvitie gornyh territorij no. 4(10) 2011 5 p.

3. Sokolova E.A., Miroshnikov A.S., Razrabotka programmnyh produktov dlja konvertacii mul''timedijnyh izobrazhenij // no. 11(38) nauchnyj zhurnal «Perspektivy nauki» 2012.

4. Chernenko V.V. Ispol'zovanie avtomatizirovannyh informacionnyh sistem v jekspozicii p. 88 // Muzei Moskvy i muzeologija XX veka: Tezisy nauchnoj konferencii (M., 25-26 nojabrja 1997 g.) // Otv. red. Ju.U. Gural'nik M.: RGGU, 1997.

5. Jel'zasser M.Je. Knjazeva N.A. Novoe izmerenie partnerstva: virtual'naja vystavka i real'noe sotrudnichestvo. // Jelektronnyj potencial muzeja: stimuly i ogranichenija, dostizhenija i problemy. Tezisy dokladov XXX Mezhdunarodnoj konferencii CIDOC-ADIT-2003. Sankt-Peterburg, 2003 pp. 55–56.

Рецензенты:

Петров Ю.С., д.т.н., профессор, зав. кафедрой «Теоретические основы электротехники», ФГБОУ ВПО «Северо-Кавказский горно-металлургический институт (Государственный технологический университет)», г. Владикавказ;

Гроппен В.О., д.т.н., профессор, зав. кафедрой «Автоматизированная обработка информации», ФГБОУ ВПО «Северо-Кавказский горно-металлургический институт (Государственный технологический университет)», г. Владикавказ.

Работа поступила в редакцию 19.07.2013.