УДК 004.652

МЕРА ИНФОРМАЦИОННОГО ПОДОБИЯ ДЛЯ АНАЛИЗА СЛАБОСТРУКТУРИРОВАННОЙ ИНФОРМАЦИИ

Бутакова М.А.[1], Климанская Е.В.[1], Янц В.И.[2]

[1]ФГБОУ ВПО «Ростовский государственный университет путей сообщения», Ростов-на-Дону, Россия (344038, Ростов-на-Дону, пл. им. Ростовского Стрелкового полка Народного Ополчения, 2), e-mail: inf-rgups@yandex.ru
[2]ФГБОУ ВПО «Ростовский государственный строительный университет», Ростов-на-Дону, Россия (344022, Ростов-на-Дону, ул. Социалистическая, 162)

В статье предложена новая мера информационного подобия для анализа слабоструктурированных документов, базирующаяся на интерференционно-волновом подходе. Дано описание предметной области исследований слабоструктурированных данных. Приведены примеры возникновения явления слабой структурированности документов. Представлены принципы организации хранения слабоструктурированных документов в базах данных и описан существующий инструментарий бессхемных баз данных и баз данных с изменяемой схемой данных. Подробно изложен принцип построения интерференционной волны и векторов интерференции. Приведена формула вычисления меры на основе векторов интерференции. Описан процесс индексации и нахождения релевантной информации по мере. Предложена модификация интерференционно-волновой меры информационного подобия в обобщенном виде. Проведено тестирование модели на экспериментальной базе данных. Установлено, что алгоритм вычислений по предложенной мере имеет линейную сложность вычислений Сделаны выводы о возможности применения предложенного метода в больших базах данных.

Ключевые слова: слабоструктурированные данные, мера подобия, поиск, индексация, базы данных

A MEASURE OF INFORMATION SIMILARITY FOR SEMISTRUCTURED INFORMATION ANALYSIS

Butakova M.A.[1], Klimanskaya E.V.[1], Yants V.I.[2]

[1] Rostov State Transport University, Rostov-on-Don, Russia, (344038, Rostov-on-Don, square n.a. Rostovskogo Strelkovogo polka Narodnogo Opolchenija, 2), e-mail: inf-rgups@yandex.ru
[2] Rostov State Building University, *Rostov-on-Don, Russia, (344022, Rostov-on-Don, street Sotsialisticheskaja, 162)*

The paper proposes a new measure of information for the analysis of similarity of semistructured documents based on interference-wave approach. A description of the subject area of research semistructured data is done. There examples of the phenomenon of weak structured documents are presented. The principles storage organization semistructured documents in databases and tools in schema-less existing databases and databases with variable data schema are described. The principle of interference wave vectors and interference is expounded. The formula calculating the measures on the basis vectors of interference is done. The process of indexing and finding relevant information on the measure is described. A modification of the interference-wave measures of similarity information in summary form is developed. Testing of the model on an experimental database is executed. Found that the proposed algorithm for computing least has a linear computational complexity conclusions about the possibility of applying the method in large databases.

Keywords: semistructured data, database, measures of similarity, search, indexing

Введение

Рост популярности *web*-служб предоставил в распоряжение пользователей различные виды информации: текст, изображение, видео, аудио, каждый из которых, довольно однороден. Вследствие этой особенности, а также информационных потребностей пользователя, информация, которая возвращается к пользователю, представлена в виде отдельных документов, что не всегда удовлетворяет ожиданиям пользователя. В действительности, доступ к информации посредством Интернет-приложений стал обычным

явлением, однако, не всегда пользователи точно и формализовано описывает свои запросы на получение информации, а, приложения дают ответ сообразно ожиданиям пользователей. Такая ситуация возникает вследствие наложения многих факторов: недостаточная выразительность языков исполнения запросов к данным, наличие многих источников информации с перекрестными ссылками, сверхбольшие объемы информации, временная недоступность, либо, перегруженность при доступе к базам данных (БД) и другие факторы. В целом, с точки зрения современных БД хранить информацию под воздействием перечисленных факторов можно, или, применяя избыточность (количественную, пространственную), или, найти подходы к описанию явления в слабоструктурированном виде. В связи с этим, развитии направлений теории БД одним из актуальных вопросов сейчас является улучшение моделей извлечения информации, путем поиска новых мер подобия и новых подходов в работе со слабоструктурированными данными. Это позволит принимать во внимание содержание и структуру слабоструктурированных документов для повышения точности результатов в соответствии с запросами пользователей Интернет-приложений, использующих БД.

Слабоструктурированные данные: реализация БД и измерение информационного подобия

Слабоструктурированные данные [6] являются формой организации данных, при которой структура документа не может быть задана заранее, а БД, хранящая такие документы допускает недоопределенности в схеме описания, а также может изменяться в течении эксплуатации. Эта форма данных содержит теги и другие маркеры для отделения семантических элементов и для обеспечения иерархической структуры записей и полей в наборе данных. В слабоструктурированных данных, сущности, принадлежащие одному и тому же классу данных, могут иметь разные атрибуты. Вопросы, касающиеся слабоструктурированных данных, их классификации и возможных путей моделирования рассматривались в работах [1,2,3].

Элементы реализации идеи слабоструктурированной обработки и хранения данных имеются в бессхемных БД, относящихся к типу NoSQL [4] систем. Их особенностью, в частности является горизонтальное масштабирование хранилища данных и поддержка поиска и индексирования по произвольным полям, а в некоторых БД имеется возможность составления произвольных запросов выборки данных. Наиболее простым способом реализации слабоструктурированного хранения данных является динамическое хранилище ключей и значений, как реализовано в БД *Redis* и *Riak*. Другим подходом к обеспечению возможности динамического изменения структуры БД является столбцовая реализация хранения данных (противоположно строковой в реляционных базах данных), при которой

есть возможность определения разного количества столбцов для различных строк. По такому принципу устроены БД *HBase, Cassandra, HyperTable*.

Особый интерес в отношении хранения и поиска представляют слабоструктурированные данные, представленные в виде цельных документов, относящихся к какой-либо категории или классу. Типичным примером слабоструктурированного документа, однако, четко попадающего во вполне определенный класс, является «счет-фактура». Он является наиболее проверяемым документом при аудите, но наибольшее количество судебных разбирательств связано именно с ним, потому, что Налоговый кодекс РФ в статье 169 «Счет-фактура» лишь общие требования по их заполнению. Таким образом, организации могут различным образом формировать нумерацию, адрес, подписи и другие реквизиты документа в процессе своей деятельности, реформирования организации и принципов учета. Счета-фактуры компаний с иностранным капиталом (инвойсы) могут содержать поля практически не используемые в отечественных документах, например, поле *Value Added Tax* (налог на потребление, являющийся некоторым аналогом отечественного налога на добавленную стоимость) Следовательно, самым гибким путем адаптации автоматизированных систем обработки таких документов является применение документо-ориентированных БД.

К таковым, в частности, относятся БД *MongoDB, CouchDB*. В БД *MongoDB* документ будет относиться и храниться в какой-либо коллекции, а форматом хранения служит структура на языке *JSON*. Схема БД MongoDB полностью изменяемая, использующая технологию *Google MapReduce*, допускающая построения широкого спектра индексов документов: уникальных, составных, геопространственных и вложенных. Для устойчивости и надежности хранения данных применяется атомарность операций, журналирование, технология асинхронной репликации с сегментацией по нескольким наборам реплик. Бессхемная БД *CouchDB* также использует массово-списочные функции *map*/*reduce*, интерфейс *REST API* для непосредственной обработки удаленных данных через *HTTP* протокол, а также формат описания данных *JSON*.

Наряду с привлекательными возможностями документо-ориентированного хранения данных есть и особенности, которые не особенно приветствуются разработчиками автоматизированных систем на основе БД. Среди таковых: отсутствие транзакций, невозможность автоматического приведения типов данных, затрудненная работа с массивами данных в отношении их сортировки и фильтрации, требование приведения данных к определенному формату и отсутствие других привычных функций БД.

Является достаточно очевидным факт, что для поиска документов в документо-ориентированных базах данных методы, которые используются в БД класса SQL, являются

малопригодными. Отметим также, что в бессхемных БД не менее важен не только точный поиск и индексация, но и поиск релевантных искомому документов. В связи с этим, важна разработка формальных методов и вычислений *меры информационного подобия*, схожести документов в смысле их информационной направленности и принадлежности к некоторому классу. В общем смысле, существует значительное количество мер в различных областях науки, начиная от простой декартовой до многогранных сложных вероятностных мер. Очевидно также, что большинство из известных мер невозможно применить в рассматриваемой области исследований.

В рамках настоящей работы ограничимся одним из перспективных подходов [7] к составлению меры информационного подобия – интерференционно-волновом. Такая мера подобия используется для получения релевантных документов в базе слабоструктурированных данных. Она показывает уровень соответствия между наиболее релевантным документом и документами БД. Мера принимает во внимание соседство лексических единиц (терминов) и тегов. Каждый из документов, принадлежащий к документам рассматриваемой БД и слабоструктурированный документ, предназначенный для поиска, должны быть предварительно обработаны с помощью индексации [5]. Принимаются во внимание два типа информации: структурная и текстовая. Структурная информация извлекается путем перехода к одному из методов представления. В данном случае, это составление карты путей документа. Текстовая информация представляет собой термин из словаря базы данных, который находится по тому или иному пути. Для генерации существенных признаков, при поиске требуемого слабоструктурированного документа и документов в БД используются оба типа информации. Оба типа информации накладываются друг на друга. В качестве результата мы получаем волновую интерференцию, которую рассмотрим подробнее. Поиск документа в БД основан на генерации волновой интерференции, посредством сравнения слабоструктурированного документа с каждым документом БД. Данное сравнение позволяет вычислить уровень подобия.

Пусть D, Q два слабоструктурированных документа, U – набор лингвистических единиц, который включает в себя документ D, и T – общее количество слов в словаре. Пусть u – лингвистическая единица, прошедшая фильтр стоп-слов и нормализованная при помощи правил и словаря лемм. Определим функцию f релевантности:

$y(u) = 3$ – лингвистическая единица u принадлежащая документу Q не существует в документе базы данных D;

$y(u) = 2$ – лингвистическая единица u принадлежащая документу Q существует в документе базы данных D, но является изолированной, без соседних слов в общих и не общих путях;

$y(u) = 1$ – лингвистическая единица u, принадлежащая документу Q существует в с как минимум 1 соседним элементом в общих и не общих путях, документа D базы данных.

Посредством интерференции волн, можно определить 3 вектора следующим образом: V_0 (а так же V_1 и V_2), представим вектор содержащий последовательности различных уровней релевантности документа, таких как $V_0[J]$ (и $V_1[J]$, $V_2[J]$, соответственно), где J весовой коэффициент последовательности 0 (1,2 соответственно). Например, векторы интерференции для волновой интерференции со значениями $V_0[0]=[0,0,1]$; $V_1[2]=[0,0,2]$ означают, что на уровне 0 искомая лингвистическая единица существует в документе, а на уровне 1 является изолированной без соседних слов в общих путях поиска документов. По аналогии с физической массой вещества, будем измерять вклад в информационную похожесть (релевантность документа) в «условных граммах», а весовые коэффициенты *J-граммами*.

Основной результат: модифицированная интерференционно-волновая мера информационного подобия

Задача меры подобия зафиксировать величину информационного подобия между обрабатываемым документом Q с каждым документом D базы данных.

Информационную меру подобия, аналогично работе [8] определим тремя векторами интерференции V_0, V_1, V_2 где каждый связан с коэффициентом, отражающим релевантность данных, найденных на каждом уровне:

$$\varpi = \frac{3\sum_{J=1}^{n}\frac{1}{\alpha_J}V_0 + 2\sum_{J=1}^{m}\frac{1}{\alpha_J}V_1 + \sum_{J=1}^{k}\frac{1}{\alpha_J}V_3}{3}\cdot 100, \qquad (1)$$

где:

$\alpha_J = T / J$ - максимальное число J-грамм в документе D и T число лингвистических единиц в документе;

n, m, k – размеры векторов V_0, V_1, V_2, соответственно.

Обобщим выражение (1) с целью нормирования меры.

$$\varpi = \left|1 - \frac{1}{d}\left(n\sum_{J=1}^{n}\frac{1}{\alpha_J}V_0 + m\sum_{J=1}^{m}\frac{1}{\alpha_J}V_1 + k\sum_{J=1}^{k}\frac{1}{\alpha_J}V_3\right)\right|, \qquad (2)$$

где $d = \max(n, m, k)$.

Чтобы оценить производительность предложенного метода, в качестве критерия было выбрано время выполнения операций. Эксперименты проводились на тестовой базе данных содержавшей 150 слабоструктурированных документов. Общий объем данных в БД – 51 *Мб*, в базе данных содержатся УМК 25 различных дисциплин, которые характеризуются неоднородным размером. Задача системы определить, к какой из дисциплин относится документ из запроса. Тестирование времени индексации и определения информационного

подобия по формуле (2), показало, что время индексации в зависимости о числа терминов в БД растет линейно, а не экспоненциально, что позволяет использовать предложенное выражение совместно и индексацией. Это делает возможным измерять информационное подобие документов в БД большого размера. Рост времени поиска, связан с увеличением размера документов базы данных, что приводит автоматически к увеличению числа слагаемых в слабоструктурированных документах в базе данных.

Заключение

В данной статье была представлена новая мера подобия, специализированная для слабоструктурированной информации. Мера состоит из понятия о волновой интерференции, извлечения векторов интерференции и затем, применения функции вычисления меры подобия. Метод основан на двух типах информации: структурной и текстовой. Структурная информация представляет собой путь из тегов ведущий к текстовой информации, представленной словарем и связанной с соседними элементами. Используя соседние элементы, имеется увеличение эффективности представления результатов поиска в БД. Система поиска в слабоструктурированной информации состоит из двух фаз: индексации и поиска. Система была проверена на тестовой базе данных и было выявлено, что время исполнения операция находится в линейной зависимости от количества терминов в БД.

Работа выполнена при финансовой поддержке РФФИ, проекты: 12-07-13120-офи_м_РЖД, 12-08-00798-а, 13-01-325-а, 13-01-00637-а, 13-08-12151-а.

Список литературы

1. Бутакова М.А. Организация хранения и обработки слабоструктурированных документов в информационно-управляющих системах на железнодорожном транспорте / Бутакова М.А., Климанская Е.В., Янц В.И. // Вестник Ростовского государственного университета путей сообщения. – 2013. – №4. – С. 42-47.

2. Климанская Е.В. Методы обработки слабоструктурированных данных в автоматизированных системах на железнодорожном транспорте / Климанская Е.В. Чернов А.В., Янц В.И. // Известия высших учебных заведений. Северо-Кавказский регион. Серия технические науки. – 2013. – №1. – С. 118-123.

3. Параценко И.Г. Классификация моделей надежности программного обеспечения / Параценко И.Г., Чернов А.В. [Электронный ресурс] // «Инженерный вестник Дона». – 2012. – №4 (часть 2). URL: http://ivdon.ru/magazine/archive/n4p2y2012/1319 (дата обращения: 05.12.2013).

4. Редмонд Э. Семь баз данных за семь недель. Введение в современные базы данных и идеологию *NoSQL* / Редмонд Э., Уилсон Д.Р. – М.: ДМК Пресс, 2013. – 384 с.
5. Bounhas I. A hierarchical approach for semi-structured document indexing and terminology extraction / Bounhas I., Slimani Y. // International Conference on Information Retrieval and Knowledge Management (CAMP). – 2010. PP. 315-320.
6. Buneman P. Semistructured data // In Proceedings of the Sixth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Tucson. – 1997. – P. 117-121.
7. Guezouli L. Gestion de documents plurimedia et recherche d'informations dans un système collaboratif, PhD Th.: Université Denis Diderot, Paris VII / AdVestigo. – 2006.
8. Guezouli L. CASIT: Content based identification of textual information in a large database / Guezouli L., Essafi H. // IEEE 24th International Conference on Advanced Information Networking and Applications Workshops. – 2010. –PP.621-625.

Рецензенты:

Ляпин А.А., д.ф.-м.н., профессор, заведующий кафедрой информационных систем в строительстве Ростовского государственного строительного университета, г.Ростов-на-Дону.

Чернов А. В., д.т.н., профессор, заведующий кафедрой прикладной математики и вычислительной техники Ростовского государственного строительного университета, г.Ростов-на-Дону.