УДК 681.3

А.М. Наместников, А.А. Филиппов

РЕАЛИЗАЦИЯ СИСТЕМЫ КЛАСТЕРИЗАЦИИ КОНЦЕПТУАЛЬНЫХ ИНДЕКСОВ ПРОЕКТНЫХ ДОКУМЕНТОВ[1]

Наместников Алексей Михайлович, *кандидат технических наук, доцент, окончил радиотехнический факультет Ульяновского государственного технического университета. Доцент кафедры «Информационные системы» факультета информационных систем и технологий УлГТУ. Имеет статьи и монографию в области интеллектуальных систем хранения и обработки информации. [e-mail: nam@ulstu.ru].*

Филиппов Алексей Александрович, *аспирант кафедры «Информационные системы» УлГТУ, окончил факультет информационных систем и технологий УлГТУ. Имеет статьи в области интеллектуальных систем хранения и обработки информации. [e-mail: al.filippov@ulstu.ru].*

Аннотация

В статье содержится описание реализации системы кластеризации проектных документов (ПД). Выполняется предварительное индексирование документов с целью получения концептуальных индексов. Онтология предметной области представляется в виде RDF-троек «объект-субъект-предикат». Для хранения концептуальных индексов применяется формат XML. В качестве репозитория используется XML-сервер Tamino.

Ключевые слова: интеллектуальная система, кластеризация, онтология, концептуальный индекс.

Alexey Mikhailovich Namestnikov, *Candidate of Engineering, Associate Professor; graduated from the Faculty of Radio-Engineering of Ulyanovsk State Technical University; Associate Professor at the Chair 'Information Systems' of the Faculty of Information Systems and Technologies; author of articles and a monograph in the field of intelligent systems for storage and processing of information. e-mail: nam@ulstu.ru.*

Alexey Alexanderovich Philippov, *post-graduate student at the Chair 'Information Systems'; graduated from the Faculty of Information Systems and Technologies of Ulyanovsk State Technical University; author of articles in the field of intellectual systems for storage and processing of information. e-mail: al.filippov@ulstu.ru.*

Abstract

The article contains a description for the implementation of a clustering system of design documents. The documents are preliminary indexed in order to obtain conceptual indexes. The domain ontology is represented in the form of RDF-triplets 'object-subject-predicate'. The XML-format is applied to store conceptual indexes. The Tamino XML-server is used as a repository.

Key words: intellectual system, clustering, ontology, conceptual index.

Введение

В настоящее время во многих проектных организациях осуществляется перевод архива проектных документов в электронный формат. В связи с этим появилась необходимость в систематизации и автоматизации работы с полученными ПД. Помочь решить данную задачу может интеллектуальный проектный репозиторий (ИПР), который способен выполнять процедуру кластеризации, обрабатывая ПД на уровне понятий предметной области, определенных в онтологии. Основными функциями реализованной интеллектуальной системы являются:

- хранение ПД в структурированном виде,
- концептуальное индексирование документов,
- концептуальная кластеризация документов.

1 Структурная схема проектного репозитория

На рисунке 1 представлено структурное решение интеллектуального проектного репозитория.

Как видно из рисунка, ИПР состоит из:

- редактора онтологий,
- подсистемы концептуальной индексации,
- концептуального кластеризатора,
- Web-фреймворка Sesame (хранилище онтологий),
- XML-сервера хранения данных Tamino.

Далее более подробно рассмотрим работу подсистемы концептуальной кластеризации, предварительно раскрыв особенности реализации редактора онтологий.

1 Работа выполнена при финансовой поддержке РФФИ, проект №10-07-00064-а.

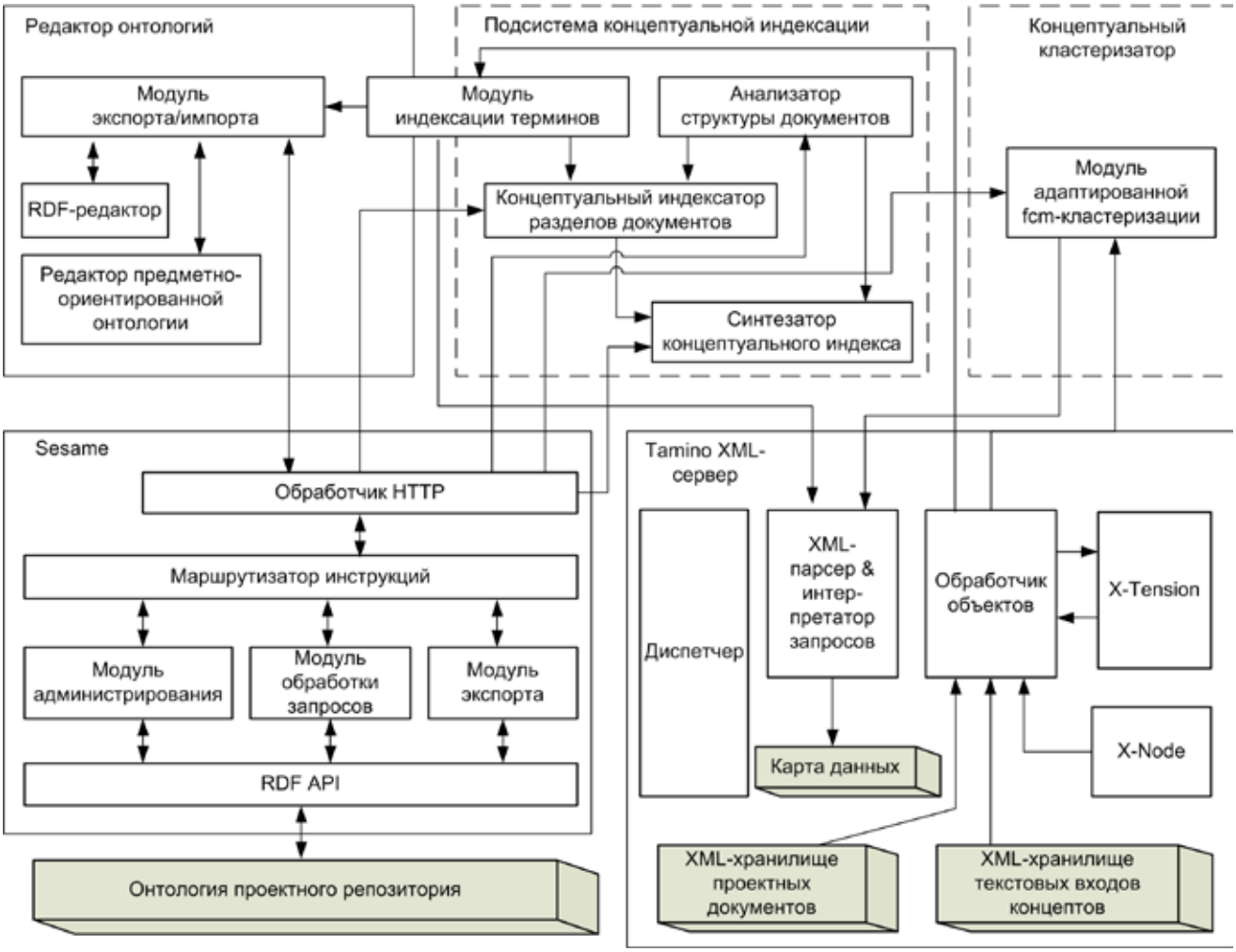

Рис. 1. Структура интеллектуального проектного репозитория

2 Редактор онтологий

Рассмотрим определение понятия онтология, представленное в работе [1]. Онтология – артефакт, структура, описывающая значения элементов некоторой системы. Неформально, онтология представляет собой некоторое описание взгляда на мир применительно к конкретной области интересов. Это описание состоит из терминов и правил использования этих терминов, ограничивающих их значения в рамках конкретной области.

На формальном уровне, онтология – это система, состоящая из набора понятий и набора утверждений об этих понятиях, на основе которых можно строить классы, объекты, отношения, функции и теории.

Основными компонентами онтологии являются:

- классы или понятия,
- отношения,
- функции,
- аксиомы,
- примеры.

Для построения онтологии интеллектуального проектного репозитория было выбрано хранилище Sesame – открытая (open source) база данных RDF с поддержкой логического вывода по RDF-тройкам и запросов на языках SeRQL и SPARQL. Оно предлагает разработчикам большой набор инструментов для работы с RDF и RDF Schema. Для взаимодействия с Sesame был разработан редактор онтологий, который призван помочь эксперту в построении онтологии предметной области. Процесс формирования онтологии представлен на рисунке 2.

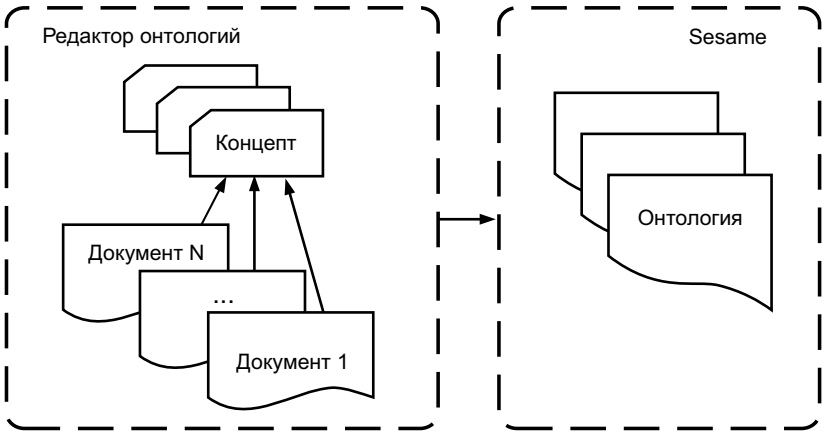

Рис. 2. Процесс формирования онтологии

Каждое понятие в предметной области описывается с помощью набора документов. Для набора терминов каждого документа применяются операции удаления стоп-слов (слов, не представляющих смысловой ценности), стемминга (процесса нахождения основы слова для заданного исходного слова, получения термов) и подсчет относительный частоты встречаемости каждого терма. На основе полученных данных формируется онтология в формате RDF, затем данный файл сохраняется на сервер Sesame.

Рассмотрим пример части онтологии ИПР. В онтологии представлены описания понятий, терминов, из которых состоят понятия, и их отношения.

Описание понятия «система»:

<Concept rdf:ID="Система"/>.

Описание термина «множество»:

<Term rdf:ID="множество"/>.

Описание отношения между понятиями (понятие «информационная система» является подпонятием понятия «система»):

```
<Concept rdf:ID="Информационная система">
    <IsASubconcept rdf:resource="Система" />
</Concept>
```

Описание отношения между понятием и термином (термин «множество» содержится в понятии «система» с частотой встречаемости 0,405465):

```
<ConceptTerm>
    <AssociatedWithConcept rdf:resource="Система" />
    <AssociatedWithTerm rdf:resource="множество" />
    <HasAFreq rdf:datatype="xsd;float">0,405465
    </HasAFreq>
</ConceptTerm >.
```

3 Подсистема концептуальной индексации

Этапы индексации проектных документов:

- загрузка документов,
- анализ структуры документов,
- удаление стоп-слов,
- стемминг,
- подсчет относительной частоты встречаемости термов,
- расчет степени выраженности понятий,
- построение концептуальных индексов для разделов и документов.

В качестве входных данных подсистемы концептуальной индексации выступают ПД. Также следует подчеркнуть, что основной единицей системы является не сам документ, а каждый его раздел в отдельности.

В самом начале процесса индексации из полученного XML-документа удаляются стоп-слова. Далее к обработанному документу применяется процесс стемминга. В рассматриваемой системе применяется стеммер Snowball. В результате данного шага получается документ, состоящий из термов. Под термом стоит понимать лексическую единицу, полученную в результате процесса стемминга.

Следующим шагом идет подсчет частоты встречаемости терма. Результирующие данные представлены в виде пары «терм-частота». После выполнения этих шагов полученный XML-документ загружается на сервер Tamino. На основе частотного портрета документа, полученного на предыдущем шаге, проводится процесс вычисления степеней выраженности понятий в разделе документа и формируется концептуальный индекс данного раздела. Затем формируется концептуальный индекс всего документа, при этом результирующая степень выраженности понятия есть дизъюнкция исходных степеней.

На рисунке 3 представлена структура подсистемы концептуальной индексации.

Математическая модель формирования концептуального индекса представлена подробно в работе [2].

Ниже представлен пример полученного концептуального индекса проектного документа в формате XML:

```
<graph>
    <vertex name="Система"
    value="0.536229" parent="ROOT" />
    <vertex name="Информационная_система"
    value="0.620224" parent="Система" />
    <vertex name="Техническая_система"
    value="0.641499" parent="Система" />
    <vertex name=
    "Медицинская_информационная_система"
```

Рис. 3. Структура подсистемы концептуальной индексации

value="0.334437" parent=
"Информационная_система" />
<vertex name="Геоинформационная_система"
value="0.403800" parent=
"Информационная_система" />
</graph>.

Таким образом, получаем, что проектный документ в ИПР представляется не в лексическом пространстве терминов, которые удается выделить в документе, а в пространстве понятий предметной области, которые зафиксированы в онтологии ИПР.

4 Реализация процесса концептуальной кластеризации

Для кластеризации концептуальных индексов в ИПР используется алгоритм кластеризации Fuzzy-c-means – метод кластеризации, который позволяет одному объекту принадлежать двум или более кластерам с определенной степенью. Этот метод (разработанный J.C. Dunn в 1973 году и улучшенный J.C. Bezdek в 1981 году) часто используется при решении задачи распознавания образов. Алгоритм основан на минимизации следующей целевой функции:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^m \left\| cI_i - cI_j^c \right\|^2,\ 1 \le m < \infty,$$

где N – количество концептуальных индексов для кластеризации;

C – количество кластеров;

m – любое действительное число больше 1;

u_{ij} – степень принадлежности концептуального индекса cI_i кластеру j;

cI_i – i-й концептуальный индекс;

cI_j^c – центр j-го кластера;

||*|| – нормализованное расстояние между концептуальным индексом и центром кластера.

Так как структура для всего набора концептуальных индексов одинакова (являются нечеткими вершинными подграфами одного и того же графа – онтологии ИПР), будем рассматривать нечеткий граф концептуального индекса, как вектор, содержащий значения понятий и их степеней выраженности [3, 4].

FCM-алгоритм состоит из следующих шагов:

Шаг 1. Инициализация

Задаются параметры кластеризации и инициализируется первоначальная матрица принадлежности концептуальных индексов кластерам $U = \left[u_{ij} \right]$.

Шаг 2. Вычисление центров кластеров

Вычисляется новое значение центров кластеров:

$$cI_j^c = \frac{\sum_{i=1}^{N} u_{ij}^m * cI_i}{\sum_{i=1}^{N} u_{ij}^m}.$$

Шаг 3. Формирование новой матрицы принадлежности

Формируется новая матрица принадлежности с учетом вычисленных на предыдущем шаге центров кластеров:

$$u_{ij} = \frac{1}{\sum_{l=1}^{C} \left(\frac{\left\| cI_i - cI_j^c \right\|}{\left\| cI_i - cI_l^c \right\|} \right)^{\frac{2}{m-1}}},$$

где u_{ij} – степень принадлежности i-го концептуального индекса кластеру j;

cI_j^c – концептуальный индекс центра j-го кластера;

cI_l^c – концептуальный индекс центра l-го кластера.

Шаг 4. Вычисление целевой функции

Вычисляется значение целевой функции, и полученное значение сравнивается со значением на предыдущей итерации.

Если разность не превышает заданного в параметрах кластеризации порогового значения, считаем, что кластеризация завершена. В противном случае переходим ко второму шагу алгоритма.

Для определения расстояния между содержимым ПД в ИПР необходимо измерить степень близости, похожести между концептуальными индексами ПД.

Будем рассматривать концептуальный индекс ПД как иерархию. Тем самым, расстояние между содержимым ПД находится через сложность превращения одной иерархии в другую, путем вычисления разности между степенями выраженности понятий, имеющих одинаковые метки (имя) [1].

Рассмотрим определение понятия «иерархия», представленное в работе [5]. Обозначим через W конечное множество объектов, $W = w_1, w_2, ..., w_l, ..., w_q$, а через H – множество непустых частей множества W, называемых таксонами и обозначаемых через h.

Иерархией H множества W называется множество подмножеств W таких, что:

- $\forall w \in W \{w\} \in H$ (терминальные вершины (листья) – одноэлементные множества);
- $W \in H$ (наибольший таксон (корень) содержит все элементы W);
- для любых вершин $h, h' \in H$ мы имеем либо $h \cap h' = 0$, либо $h \subset h'$, либо $h' \subset h$.

Таким образом, иерархия – это многоуровневая структура, в которой объекты, находящиеся в одном таксоне на некотором j-м уровне, остаются в одном таксоне на $(j+1)$-м и всех других более высоких уровнях.

Первому уровню соответствуют терминальные вершины (п. 1 в определении иерархии), а последнему, максимальному, уровню (обозначим его через m) – наибольший таксон, содержащий все элементы W; этот таксон можно обозначить тем же символом W (п. 2 в определении иерархии). На каждом уровне происходит или не происходит объединение таксонов (п. 3 в определении иерархии).

Обозначим точкой каждый таксон иерархии. Тогда вершины иерархии будут описываться именем понятия c_{id} и его степенью выраженности в ПД f_{id}. Для примера рассмотрим нахождение расстояния между двумя документами, представленными иерархиями H^1 и H^2 соответственно.

$$H^1\begin{cases} c_{111}=1(f_{111}=0,2);\\ c_{211}=11(f_{211}=0,6);\\ c_{221}=12(f_{221}=0);\\ c_{311}=111(f_{311}=0,1);\\ c_{321}=112(f_{321}=0,4); \end{cases}$$

$$H^2\begin{cases} c_{112}=1(f_{112}=0);\\ c_{212}=11(f_{212}=0,3);\\ c_{222}=12(f_{222}=0,4);\\ c_{312}=111(f_{312}=0,5);\\ c_{322}=112(f_{322}=0). \end{cases}$$

Редакционное расстояние определяется на основе вычисления стоимости редакционной операции:

$$v(lj1,\ lj2)=f_{lj1}-f_{lj2},$$

используя модифицированный алгоритм схожих пар. При этом рассматриваются схожие вершины h_{lj1} и h_{lj2} с учетом соответствия меток, т. е. при $c_{lj1}=c_{lj2}$.

Рассмотрим пример нахождения расстояния между концептуальными индексами, представленный на рисунке 4.

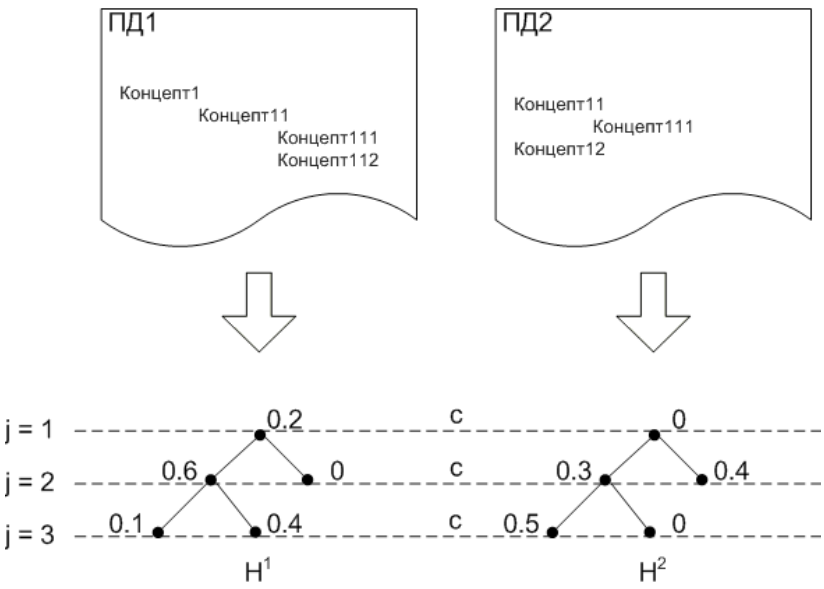

Рис. 4. Пример иерархий двух концептуальных индексов

Рассчитаем расстояние τ^* между иерархиями H^1 и H^2 двух ПД (рис. 4) с помощью модифицированного алгоритма схожих пар:

$$\tau^*=(0,2-0)+((0,6-0,3)+(0,4-0))+((0,5-0,1)+(0,4-0))=0,2+0,7+0,8=1,7.$$

Коэффициент нормализации T рассчитаем из расстояния между иерархиями, учитывая, что метки вершин не совпадают:

$$\forall l: c_{lj1}\neq c_{lj2}.$$

Таким образом имеем:

$$T=(0,2+0)+(0,6+0,3+0+0,4)+(0,1+0,5+0,4+0)=2,5.$$

Найдем расстояние между содержимым ПД с учетом их концептуальных индексов:

$$\tau=\frac{\tau^*}{T}=\frac{1,7}{2,5}=0,68.$$

Заключение

Предлагаемый в работе метод кластеризации позволяет работать с проектными документами репозитория не на уровне терминов, встречающихся в текстах, а на уровне понятий предметной области, набор которых зафиксирован в онтологии. Зависимость результатов кластеризации от описания предметной области позволяет управлять процессом кластеризации, получая различные классификационные схемы содержимого репозитория, основываясь на активном подмножестве понятий онтологии.

СПИСОК ЛИТЕРАТУРЫ

1. Онтологии и тезаурусы: учеб. пособие / Соловьев В. Д. [и др.]. – М., 2006. – 157 с.

2. Наместников А.М., Филиппов А.А. Концептуальная индексация проектных документов // Автоматизация процессов управления. – 2010. – № 2 (20). – С. 34–39.

3. Наместников А.М., Филиппов А.А. Нечеткая кластеризация концептуальных индексов проектных документов // Интегрированные модели и мягкие вычисления в искусственном интеллекте: сб. науч. тр. 6-й Межд. науч.-тех. конф. (Коломна, 16-19 мая 2011 г.). В 2 т. Т 2. – М. : Физматлит, 2011. – С. 958–968.

4. Наместников А.М., Чекина А.В., Корунова Н.В. Интеллектуальный сетевой архив электронных информационных ресурсов // Программные продукты и системы. – 2007. – № 4. – С. 10–13.

5. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск : ИМ СО РАН, 1999. – 270 с.