ОЦЕНКА РЕЛЕВАНТНОСТИ ЭЛЕКТРОННЫХ ДОКУМЕНТОВ К ОНТОЛОГИИ В СИСТЕМЕ ИНФОРМАЦИОННОГО МЕНЕДЖМЕНТА♦

А. С. Комарова

Санкт-Петербургский институт информатики и автоматизации РАН

СПИИРАН, 14-я линия ВО, д. 39, Санкт-Петербург, 199178

<komarova@iias.spb.su>

УДК 681.3

Комарова А. С. **Оценка релевантности электронных документов к онтологии в системе информационного менеджмента** // Труды СПИИРАН. Вып. 3, т. 1. — СПб.: Наука, 2006.

Аннотация. *Приводится обзор классических оценок близости объектов и их использование для оценки близости текстов и онтологий. Рассматривается адаптация существующих методов оценки близости для случая системы доступа к электронным документам, на основе использования онтологии. Предлагается метод оценки релевантности электронных документов заданному фрагменту онтологии.* — Библ. 27 назв.

UDC 681.3

Komarova A. S. **Estimation of Relevance of Electronic Documents to Ontology in Information Management System** // SPIIRAS Proceedings. Issue 3, vol. 1. — SPb.: Nauka, 2006.

Abstract. *An overview of classical similarity measures between objects and their usage for estimation of similarity between texts and ontologies are presented. An adaptation of existent methods of similarity measures in case of ontology-driven system access to electronic documents is suggested. A method of relevance estimating of electronic documents to starting fragment of ontology is proposed.* — Bibl. 27 items.

1. Введение

Задачей исследования данной работы являются метрики оценок релевантности полученных результатов в системах информационного менеджмента. Работа была выполнена в рамках проекта, в котором использовалась онтологическая модель [25] представления знаний проблемной области. В системах информационного менеджмента используются различные источники информации, например, базы данных, базы знаний, электронные документы, датчики, эксперты. В данной работе в качестве источников информации рассматривались электронные документы. Документ представляется как содержимое (текст) и метаданные. Для поиска документов пользователь вводит запрос на естественном языке, описывающем его задачу, и в результате получает набор документов из корпуса, соответствующий этой задаче.

В разработанном подходе сначала производится обработка запроса на естественном языке (т.е. распознавание слов, поиск незначимых слов, проверка орфографии, приведение к каноническому виду — выделение значимой части слова [15]), а затем отображение значимых слов в онтологию проблемной области. По результатам отображения формируется фрагмент онтологии, на основании которого выбираются документы из корпуса.

Современные методы оценки качества полученных результатов строятся на основе обратной связи с пользователями [11]. Полученные результаты срав-

♦Работа выполнена при финансовой поддержке РФФИ (проекты № 05-01-00151 и 06-07-89242), Президиума РАН (проект № 2.35), ОИТВС РАН (проект № 1.9) и СПбНЦ РАН (проект № 78 региональной научной программы 2006 года).

ниваются с ожидаемыми. Самой распространенной оценкой правильности является пара Precision и Recall [18, 20, 27]. Они представляются как:

$$\text{Precision} = \frac{|R \cap A|}{|A|}, \quad \text{Recall} = \frac{|R \cap A|}{|R|},$$

где R — ожидаемый результат; A — полученный результат.

Значение Recall характеризует полноту результата, а Precision точность результата. Оценки принимают значения в интервале $[0;1]$. В случае, если полученный результат полностью совпадает с ожидаемым, обе оценки принимают значение единица.

Для улучшения качества результатов используется оценка релевантности документа ожиданиям пользователя. Документы сортируются согласно оценке релевантности. С помощью этой оценки отсеиваются документы, недостаточно релевантные запросу.

Понятие релевантности не является специфичным для систем информационного поиска и систем доступа к информации. Оно появилось из философских теорий, и изучается многими направлениями науки [12].

В системах, предоставляющих доступ к информации, как и в поисковых системах, релевантность вычисляется на основе оценок близости. В настоящей работе она определяется как функция от двух оценок близости s_1 и s_2 и поправочного коэффициента k:

$$\text{Rel} = F(s_1, s_2, k),$$

где s_1 — мера близости выбранного фрагмента онтологии запросу пользователя;

s_2 — мера близости фрагмента онтологии и документов;

k — коэффициент, вычисляемый на основе метаданных документа.

Использование поправочного коэффициента k важно, когда оценки близости нескольких документов фрагменту онтологии одинаковы. Областью исследования данной работы является выбор метрики близости для вычисления близости фрагмента онтологии и документов, а также выбор набора метаданных для вычисления поправочного коэффициента.

Математически близость объектов характеризуется неотрицательным числом, которое называется расстоянием. Если это число велико, то объекты далеки друг от друга. Если оно мало, то объекты близки. Расстояние (d) должно удовлетворять следующим условиям:

- $d_{ij} \geq 0$ расстояние всегда больше либо равно нулю;
- $d_{ii} = 0$ расстояние равно нулю тогда и только тогда, когда измеряется расстояние от объекта до самого себя;
- $d_{ij} = d_{ji}$ расстояние симметрично;
- $d_{ij} \leq d_{ik} + d_{kj}$ расстояние удовлетворяет неравенству треугольника (Рис. 1).

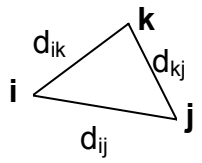

Рис. 1. Объекты и расстояния между ними.

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Метрикой близости называется только та величина, которая удовлетворяет всем указанным условиям одновременно [4, 7, 10, 13].

Расстояние является оценкой удаленности, так как оно возрастает соответственно удалению объектов друг от друга. Для того чтобы перейти к измерению близости объектов с помощью расстояния выполняется следующее действие:

– значение расстояния нормализуется, и обратная величина, являющаяся оценкой близости, максимизируется;
– или минимизируется значение расстояния.

Оценка близости объектов может быть установлена экспертом в данной области. Но этот способ содержит ряд недостатков (например, мнения экспертов, принимающих решение могут быть различными, временные затраты на подобную работу могут быть значительными). Поэтому задача автоматической оценки близости объектов является актуальной.

В данной работе представлены оценки, в которых в качестве сравниваемых объектов выступают онтология (часть онтологии) и документ. Для построения критериев близости использованы следующие элементы онтологии: названия классов (как текст), названия атрибутов (как текст), связи между классами и связей принадлежащих атрибутам классам [25] (например, как удалены классы друг от друга); а из документов: слова (как текст) и связи между ними[1].

В разделе 2 представлено исследование классических оценок близости объектов. В разделе 3 и в разделе 4 показано как классические методы применяются для оценки близости документов и для оценки близости онтологий. В разделе 5 показано, как для оценки близости между фрагментом онтологии и документом можно адаптировать классические методы, используя оценки для текстов и для онтологий. В разделе 6 приводится анализ возможности использования метаданных документов для расчета поправочного коэффициента при оценке релевантности полученного результата ожидаемым. Разделе 7 подводит итог проведенной работе.

2. Классические способы оценки близость–удаленность двух объектов

В данном разделе представлен набор широко используемых метрик для оценки близости двух объектов, состоящих из набора элементов. Элемент является характеристикой некоторого свойства объекта. Задача сравнения объектов может решаться как попарное сравнение элементов одного и того же свойства из различных объектов и объединение полученных оценок.

В данном разделе используются следующие обозначения для входных данных:

- объекты i и j;
- множество из n свойств $\{b_1,...,b_n\}$;
- элементы, характеризующие свойства объектов i и j, $\{x_{i1},...,x_{in}\}$ и $\{x_{j1},...,x_{jn}\}$.

Ниже рассматриваются оценки близости s_{ij} и расстояния d_{ij} для двух видов элементов объектов — двоичных и весовых.

[1]Под текстом в работе подразумевается набор букв и символов в определенном порядке.

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

2.1. Двоичные элементы объектов

В случае использования двоичных элементов объекта элемент принимает значение единица, если данное свойство *есть* у объекта, или ноль, если данного свойства *нет*.

Оценки сравнения объектов, состоящих из двоичных элементов, основываются на значениях:

– мощность множества свойств, присутствующих в обоих объектах, т.е. $p = \left|\{b_k : b_k \in i,\ b_k \in j\}\right|$;

– мощность множества свойств, присутствующих в объекте i и отсутствующих в j, т.е. $q = \left|\{b_k : b_k \in i,\ b_k \notin j\}\right|$;

– мощность множества свойств, отсутствующих в объекте i и присутствующих в j, т.е. $r = \left|\{b_k : b_k \notin i,\ b_k \in j\}\right|$;

– мощность множества свойств, отсутствующих в обоих объектах, т.е. $t = \left|\{b_k : b_k \notin i,\ b_k \notin j\}\right|$.

Существует три распространенные оценки:

- *Коэффициент Джакарда,* оценивающий близость. Его значение увеличивается пропорционально тому, на сколько элементы близки друг другу:

$$s_{ij} = \frac{p}{p+q+r}.$$

Так же коэффициент Джакарда называют оценкой асимметричной информации [17]. Требованием к входным объектам является наличие хотя бы одного свойства.

Значение оценки является вещественное число в интервале $[0;1]$.

- *Простейшим коэффициентом соответствия* (*Simple Matching Coefficient*) называют

$$s_{ij} = \frac{p+t}{n}.$$

Данный коэффициент является мерой близости двух объектов. В отличие от коэффициента Джакарда, простейший коэффициент соответствия дает оценку симметричной информации. Результатом является вещественное число в интервале $[0;1]$.

- Третья оценка — *расстояние Хемминга* — оценивает количество различных элементов между объектами и является оценкой удаленности объектов

$$d_{ij} = q + r.$$

Результатом является натуральное число или ноль.

Оценки сравнения объектов, состоящих из двоичных элементов, распространены из-за простоты их вычисления. Однако, в случаях, когда такие оценки являются недостаточными, используются объекты, состоящие из весовых элементов.

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

2.2. Весовые элементы объектов

Элемент содержит больше информации о свойстве объекта, если он характеризует не просто наличие данного свойства, но и его значимость для объекта. Т.е. элемент содержит вес свойства объекта.

Для сравнения двух объектов веса их элементов нормируются к общей величине.

В табл. 1 представлены распространенные метрики расстояния и близости для объектов, с весовыми характеристиками свойств. В указанных формулах предполагается, что веса элементов больше либо равны нулю.

Таблица 1

Метрики расстояния и близости для объектов с весовыми элементами

№	Описание метрики	Формула	Возможный результат
1	Расстояние Евклида.	$d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2}$	$R+$
2	Расстояние Спирмена.	$d_{ij} = \sum_{k=1}^{n}(x_{ik} - x_{jk})^2$	$R+$
3	Расстояние абсолютной величины.	$d_{ij} = \sum_{k=1}^{n} \mid x_{ik} - x_{jk} \mid$	$R+$
4	Расстояние Минковски. $\lambda \in (0;+\infty)$	$d_{ij} = \sqrt[\lambda]{\sum_{k=1}^{n}(x_{ik} - x_{jk})^{\lambda}}$	$R+$
5	Канберрийское расстояние.	$d_{ij} = \sum_{k=1}^{n} \frac{\mid x_{ik} - x_{jk} \mid}{\mid x_{ik} \mid + \mid x_{jk} \mid}$	$R+$
6	Расстояние Брай Куртиса (Bray Curtis). Также используется как способ нормализации.	$d_{ij} = \frac{\sum_{k=1}^{n} \mid x_{ik} - x_{jk} \mid}{\sum_{k=1}^{n}(x_{ik} + x_{jk})}$	[0;1]
7	Угловое разнесение.	$s_{ij} = \frac{\sum_{k=1}^{n} x_{ik} \cdot x_{jk}}{(\sum_{k=1}^{n} x_{ik}^2 \cdot \sum_{r=1}^{n} x_{jr}^2)^{\frac{1}{2}}}$	[0;1]
8	Коэффициент корреляции. Аналогична угловому разнесению, но координаты приводятся к началу координат с помощью их среднего значения.	$s_{ij} = \frac{\sum_{k=1}^{n}(x_{ik} - \overline{x_i}) \cdot (x_{jk} - \overline{x_j})}{(\sum_{k=1}^{n}(x_{ik} - \overline{x_i})^2 \cdot \sum_{r=1}^{n}(x_{jk} - \overline{x_j}))^{\frac{1}{2}}}$	[0;1]

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Оценка близости углового разнесения (табл. 1) является аналогом косинуса угла между векторами, т.е. объект рассматривается как вектор, а элементы объекта — как его компоненты. В общем случае эта оценка может принимать значения в интервале $[-1;1]$, но, поскольку значения элементов объекта ограничены положительными числами, область была сужена до интервала $[0;1]$.

Для избежания неопределенности в случае, когда веса характеристики в обоих объектах равны нулю, в оценках канберрийское расстояние, расстояние Брай Куртиса и угловое разнесение (табл. 1) считается, что $\frac{0}{0} = 0$.

Когда коэффициент λ в оценке удаленности расстояние Минковски (табл. 1) принимает значение единица, то результат совпадает с результатом оценки расстояния абсолютной величины (табл. 1), значение два — с результатом оценки Евклида (табл. 1), а при значении $+\infty$ совпадает с результатом оценки, называемой расстоянием Чебышева.

3. Задача сравнения двух текстов

Классические методы оценки близости, описанных в предыдущем разделе, могут быть применены для задачи сравнения двух текстов.

Не смотря на то, что задача автоматического сравнения текстов родилась давно, она остается востребованной по причине ограниченности возможностей человека — большой текст сложно охватить памятью и, следовательно, характеризовать.

Впервые оценка близости текстов была выполнена в 1915 году Н.А. Морозовым [2]. Его целью было определение авторства на основе словоупотреблений часто встречающихся слов (например, союза «и» или предлога «в»). Каждому тексту привязывалась по некоторому правилу точка на плоскости, а близость текстов определялась расстоянием между точками.

В современных информационных системах методы сравнения текстов основываются на оценке использования тех или иных слов. Для оценки близости тестов используются классические метрики, представленные в разделе 2. В качестве объектов в них выступают тексты. Свойствами объектов для текстов являются слова, которые в них использованы.

В задаче определения авторского стиля [4] используется оценка близости всех слов текста или наиболее употребляемые в речи. В задачах вычисления близости запроса пользователя, заданного на естественном языке, документу [11, 16] оцениваются только значимые слова. Поэтому, в работах формируется список незначимых слов [15], которые исключаются из рассмотрения. Такими словами выступают, например, союзы («и», «а») и предлоги («в», «к»).

Широкая распространенность двоичных характеристик слов обусловлена сравнительной простотой при вычислении оценок. Элемент равен единице, если слово встречается в тексте, и нулю — если такого слова нет.

В случае весовых характеристик слов текста словам задаются веса по некоторому правилу. Текст рассматривается как вектор, состоящий из компонентов — весов слов.

Основной характеристикой, используемой при вычислении веса слов, является частота встречаемости слова в документе [3, 23, 24]. Другими известными характеристиками являются:

- количество уникальных слов в тексте [3, 23, 24];

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

- количество рассматриваемых текстов (для корпуса документов) [3, 23];
- количество текстов, содержащих данное слово (для корпуса документов) [3, 23];
- общее количество рассматриваемых уникальных слов [24].

В проекте [6] для каждого документа строился набор пар — слово и его вес. Для сравнения документов использовалась формула углового разнесения (табл. 1). При разработке поисковой системы, описанной в работе [3], была выбрана оценка удаленности расстояние абсолютной величины (табл. 1). При этом для нахождения близости оценки расстояния минимизировалась.

4. Задача сравнения онтологий

В настоящее время онтологии широко используются для представления знаний проблемных областей [25]. В настоящей работе свойствами онтологии являются классы, связи между классами и атрибуты. В данном разделе рассмотрены оценки близости объектов, являющихся частями онтологии. Оценки основаны на классические методах, описанных в разделе 2.

4.1. Двоичные характеристики

Бинарные характеристики свойств онтологий определяют наличия или отсутствия данного свойства у объекта.

В работах [22, 23] была предложена оценка близости для классов из разных онтологий, вычисляемая как взвешенная сумма близостей:

- множества синонимов имени каждого класса;
- свойств (структурные элементы классов, функций над этим классом, атрибутов) каждого класса;
- семантических соседей каждого класса[2].

Для каждой группы из списка использовалась оценка близости Джакарда. Отличием от классического коэффициента Джакарда являлось использование весового коэффициента в знаменателе, т.е. адаптированная метрика Джакарда представляется следующим образом:

$$s_{ij} = \frac{p}{p + \alpha_{ij} \cdot q + (1 - \alpha_{ij}) \cdot r}.$$

Весовой коэффициент α_{ij} строился на основе удаленности рассматриваемых классов от вершины онтологии.

4.2. Весовые характеристики свойств онтологии

В ряде проектов для оценки близости онтологий использовались не бинарные, а весовые характеристики.

Например, в проекте [21] была предложена весовая оценка связей внутри онтологии. Связь между классами C_j, C_k имеет вес, вычисленный по формуле:

[2]Семантическими соседями класса называются классы, удаленные от данного не более, чем на заданную в систему величину.

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

$$W(C_j, C_k) = \frac{\sum_{i=1}^{n} n_{ijk}}{\sum_{i=1}^{n} n_{ij}} * \frac{1}{\sqrt{n_k}},$$

где n — общее количество сущностей классов онтологии;

n_{ijk} — количество сущностей C_i, связанных и с сущностью C_j, и с сущностью C_k;

n_{ij} — количество сущностей C_i связанных с сущностью C_j;

n_k — количество связей того же типа, что и связь $C_j - C_i$.

Такая оценка аналогична известной частотной оценке веса слова в документе tf–idf [19].

5. Задача сравнения текста и онтологии

При сопоставлении онтологии и документов требуются метрики их близости. Задача отличается от вышеперечисленных в первую очередь тем, что объекты сравнения разнородны, так как первый объект — онтология (или фрагмент онтологии), второй — документ.

В работе предложены подходы к решению этой задачи на основании метрик близости текстов и онтологий, описанных в разделах 3 и 4.

5.1. Представление онтологии как текста

Первый способ опирается на представление элементов онтологии как текста (т.е. на использование названий классов и атрибутов). Сравнение происходит по схеме текст–текст. В случае весовых оценок формируется вектор слов онтологии, при этом словам из выбранного фрагмента онтологии назначается вес единица, а остальным словам из онтологии назначается вес ноль или вес, рассчитанный на основании связей отделяющих класс от фрагмента.

5.2. Приведение документа к онтологии

Задача автоматического построения онтологии из документа трудно решаема. Но такие решения существуют. В результате формирования онтологии из документа задача сравнения онтологии и документа сводится к задаче сравнения онтологий, рассмотренной в разделе 4.

Данный подход был использован в работе [5], где сравнение онтологий производилось в два этапа. На первом этапе вычисляется оценка удаленности пары элементов, один из которых выделен в документе, а другой в контексте исследования (выраженным фрагментом онтологии и набором ключевых слов). На втором этапе вычисляется четыре оценки близости между элементами онтологий: между классами, между связями, между сущностями и между ключевыми словами. Каждая оценка второго этапа основываются на соответствующих оценках, полученных на первом этапе, и вычисляется по следующей формуле:

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

$$s = \frac{\sum_{j \in d} \left[\sum_{i=1}^{|\mathrm{ng}(j)|} \frac{1}{d_{ij} + 1} \right]}{|d|},$$

где i — элемент онтологии документа;

j — элемент онтологии проблемной области;

d_{ij} — расстояние между выбранными элементами;

$|\mathrm{ng}(j)|$ — элементы онтологии, соседние с элементом j.

Затем вычисляется релевантность документа контексту исследования как сумма полученных оценок близости.

6. Метаданные

В данной работе документ представляется как содержимое (например, текст) и метаданные документа. Метаданные это структурные данные о данных. Они описывают содержимое, качество, условия использования и другие описательные характеристики данных [8]. Стандартизованные описательные метаданные предоставляют мощный механизм для улучшения результатов поиска документов приложениями [14].

В рассматриваемой работе метаданные позволяют определить соответствие документа исходным данным при оценке релевантности выбранного документа. На основе метаданных формируется поправочный коэффициент для вычисления релевантности документов. Использование поправочного коэффициента важно, когда оценки близости нескольких документов фрагменту онтологии одинаковы.

Для расчета поправочного коэффициента метаданным документа задаются нормализованные веса. Для всех документов матрица нормализованных значений весов метаданных:

$$N = \begin{vmatrix} v_{11} & \dots & v_{1j} & \dots & v_{1n} \\ \dots & & \dots & & \dots \\ v_{i1} & \dots & v_{ij} & \dots & v_{11} \\ \dots & & \dots & & \dots \\ v_{p1} & \dots & v_{pj} & \dots & v_{1n} \end{vmatrix},$$

где n — количество полученных документов;

p — количество метаданных;

v_{ij} — вес i-ого элемента из набора метаданных в j-ом документе.

Поправочный коэффициент для *j*-ого документа предлагается вычислять по формуле:

$$k_j = \frac{\sum_{j=1}^{p} v_{ij}}{p}.$$

Для формирования набора метаданных, используемого для данной задачи, были изучены популярные стандарты. На сегодняшний день одним из широко используемых стандартов, предоставляющий набор элементов метаданных,

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

является Dublin Core [26]. Он имеет два уровня: Simple и Qualified. В Simple Dublin Core содержится 15 элементов; в Qualified Dublin Core добавляется еще 7 добавочных элементов, которые улучшают семантическое описание ресурсов. Ниже представлены элементы стандарта [9]:

- Название (Title): Название ресурса.
- Создатель (Creator): Данные, описывающие создателя содержимое ресурса.
- Предмет (Subject): Тема содержания ресурса (слова, ключевые фразы или коды классификации, которые описывают тему ресурсы).
- Описание (Description): Содержания ресурса (оглавление, реферат).
- Издатель (Publisher): Данные, описывающие опубликование ресурса.
- Соисполнитель (Contributor): Данные, описывающие внесение обновления в содержимое ресурса.
- Дата (Date): Дата, связанная с событием из жизненного цикла ресурса (создан, действителен, доступен, выпущен, изменен).
- Тип (Type): Природа или жанр содержимого ресурса (например, Рисунок, Звук, Текст).
- Формат (Format): Физическое или цифровое представление ресурса.
- Идентификатор (Identifier): Недвусмысленная ссылка на ресурс с данным содержанием (например, Uniform Resource Identifier — URI, Digital Object Identifier — DOI, International Standard Book Number — ISBN).
- Источник (Source): Ссылка на ресурс, от которого можно получить данный ресурс.
- Язык (Language): Язык интеллектуального содержания ресурса (английский, русский, и другие).
- Отношение (Relation): Ссылка на связанный ресурс («является версией», «имеет версию», «замещен», «замещает», «требуется», «требует», «является частью», «имеет часть», «указан», «указывает», «является форматом», «имеет формат»).
- Охват (Coverage): Границы, как правило, включают пространственное расположение ресурса (название места или географические координаты), временной период (заголовок периода, дата или период даты) или власть (такую как название административной единицы).
- Права (Rights): Информация о правах, содержащаяся в и вне ресурса.

В данной работе ресурсами являются электронные документы из корпуса.

Этот список представляет набор из наиболее популярных, но не всех используемых, метаданных. Формирование набора метаданных для расчета поправочного коэффициента при оценке релевантности выбранных документов на основе использования онтологии является задачей дальнейшего исследования.

7. Заключение

В системах доступа к электронным документам одной из важнейших задач является оценка качества формируемых результатов, которая отображает релевантность выбранных документов проблемам пользователя. В ходе исследовательского проекта было предложено описывать проблему при помощи онтологии. Данная работа была посвящена исследованию классических оценок

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

близости объектов, их использования для оценки близости текстов и онтологий и адаптации их для оценки близости онтологии и документов.

В ходе дальнейших исследований предполагается выбрать набор метаданных, наиболее удобный для основанной на использовании онтологии системы доступа к электронным документам. Предполагается проведение серии экспериментов для выбора наиболее оптимальных оценок близости онтологии и документа. Будет проведено исследование функции оценки релевантности выбранных документов ожиданиям пользователя, основанной на оценках близости между фрагментом онтологии и запросом пользователя, между онтологией и документами и поправочным коэффициентом, а также исследование оценок близости между фрагментом онтологии и запросом пользователя.

Литература

1. *Кураленок И., Некрестьянов И.* Оценка систем текстового поиска // Программирование. 2002. Вып. 28, № 4. С. 226–242.
2. *Морозов Н. А.* Лингвистические спектры: средство для отличения плагиаторов от истинных произведений того или иного неизвестного автора. Стилеметрический этюд // Известия отд. русского языка и словесности Имп. Акад. наук. 1915. Т. XX, кн. 10.
3. *Трегубов А. А., Кононова Т. С.* Алгоритмические основы разработки поисковой системы // Четвертая Всероссийская конференция «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». 2002. С. 170–177.
4. *Хмелев Д.* Использование информационной метрики в анализе текстового материала на примере корпуса текстов А. и Б. Стругацких [Электронный ресурс] // <http://www.compression.ru/download/articles/classif/khmelev_2004_text_analysis_metrics.pdf> (по состоянию на 05.04.2006).
5. *Aleman-Meza B., Burns P., Eavenson M., Palaniswami D., Sheth A. P.* An Ontological Approach to the Document Access Problem of Insider Threat // IEEE International Conference on Intelligence and Security Informatics (ISI-2005). 2005. P. 486–491.
6. *Cetintemel U., Franklin M. J., Giles C. L.* Self-Adaptive User Profiles for Large-Scale Data Delivery // Proceedings of the 16th International Conference on Data Engineering. 2000. P. 622–633.
7. *Charikar M. S.* Similarity Estimation Techniques from Rounding Algorithms // Proceedings 34th Ann. ACM Symp. Theory of Computing. 2002. P. 380–388.
8. *Deng Y.* The Metadata Architecture for Data Management in Web-based Choropleth Maps [Электронный ресурс] // <http://www.cs.umd.edu/hcil/census/JavaProto/metadata.pdf> (по состоянию на 05.04.2006).
9. Dublin Core Metadata Element Set, Version 1.1: Reference Description [Электронный ресурс] // <http://purl.org/DC/documents/rec-dces-19990702.htm> (по состоянию на 05.04.2006).
10. *Egecioglu Ö., Ferhatosmanoglu H., Ogras Ü. Y.* Dimensionality Reduction and Similarity Computation by Inner-Product Approximations // IEEE Trans. Knowl. Data Eng. 2004. Vol. 16, no. 6. P. 714–726.
11. Google [Электронный ресурс] // <http://www.google.com> (по состоянию на 05.04.06).
12. *Greisdorf H.* Relevance: An Interdisciplinary and Information Science Perspective // Informing Science. 2000. Vol. 3, no. 2. P. 67–72.
13. *Guha S., Rastogi R., Shim K.* ROCK: A Robust Clustering Algorithm for Categorical Attributes // Proceedings of the 15th International Conference on Data Engineering. 1999. P. 512–521.
14. *Hillman D.* Using Dublin Core [Электронный ресурс] // <http://dublincore.org/documents/usageguide/> (по состоянию на 05.04.2006).
15. *Hinselmann T., Smirnov A., Pashkin M., Chilov N., Krizhanovsky A.* Implementation of Customer Service Management System for Corporate Knowledge Utilization // Proceedings of the 5th InternationalConference on Practical Aspects of Knowledge Management (PAKM 2004). 2004. P. 475–486.
16. k42 [Электронный ресурс] // <http://www.empolis.co.uk> (по состоянию на 05.04.06).
17. Kardi Tekmono's Page [Электронный ресурс] // <http://people.revoledu.com/kardi/tutorial/Similarity/Jaccard.html> (по состоянию на 05.04.2006).

18. *Moench E., Ullrich M., Schnurr H., Angele J.* SemanticMiner — Ontology-Based Knowledge Retrieval // Journal of Universal Computer Science (J.UCS). Special Issue of selected papers of the WM2003. 2003. Vol. 9. P. 682–696.
19. *Orasan C., Pekar V., Hasler L.* A Comparison of Summarisation Methods Based on Term Specificity Estimation // Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-04). 2004. P. 1037–1041.
20. *Peng J.* Multi-class Relevance Feedback Contentbased Image Retrieval // Computer Vision and Image Understanding. 2003. Vol. 90, no. 1. P. 42–67.
21. *Rocha C., Schwabe D., Aragao M. P*. A Hybrid Approach for Searching in the Semantic Web // Proceedings of the 13th International World Wide Web. 2004. P. 374–383.
22. *Rodríguez M. A., Egenhofer M. J.* Determining Semantic Similarity among Entity Class from Different Ontologies // IEEE Transactions on Knowledge and Data Engineering. 2003. Vol. 15, no. 2. P. 442–465.
23. *Rodríguez A., Egenhofer M.* Comparing Geospatial Entity Classes: An Asymmetric and Context-Dependent Similarity Measure // International Journal of Geographical Information Science. 2004. Vol. 18, no. 3. P. 229–256.
24. *Seo Y., Giampapa J. A., Sycara K*. Text Classification for Intelligent Portfolio Management // Technical report. Robotics Institute, Carnegie Mellon University. 2002. No. CMU-RI-TR-02-14.
25. *Smirnov A., Pashkin M., Chilov N., Levashova T., Krizhanovsky A.* Ontology-Driven Knowledge Logistics Approach as Constraint Satisfaction Problem / Eds. Meersman R., Tari Z., Schmidt D. C. et al // On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. Lecture Notes in Computer Science. 2003. Vol. 2888. P. 535–652.
26. The Dublin Core Metadata Initiative [Электронный ресурс] // <http://dublincore.org> (по состоянию на 05.04.2006).
27. *Zhai C., Cohen W. W., Lafferty J*. Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval // Proceedings of ACM SIGIR 2003. 2003. P. 10–17.