УДК 025.4.025:519.767.2

ПОСТРОЕНИЕ ТЕМАТИЧЕСКИХ СТРУКТУР ПРЕДМЕТНЫХ ОБЛАСТЕЙ

Васина Е.Н.,[1] **Козлова И.В.**[1]

[1] *ФБГОУ ВПО«РЭУ им. Г.В. Плеханова»Минобрнауки РФ, г. Москва, 117997, Стремянный пер., 36*

Приведен обзор инструментальных средств представления результатов поиска в виде классификационных схем предметных областей или различных тематических структур. Показано, что тенденции развития поисковых систем заключаются в постепенном расширении традиционных функций за счет подключения к поисковым механизмам интеллектуальных аналитических возможностей. Рассмотрена формально-математическая постановка задачи экспликации тематической структуры предметной области из множества документов, полученного в результате информационного поиска. Описываются модель и процесс построения тематической структуры на основе анализа терминологической сети и установления семантических отношений между понятиями. Терминологическая сеть является объектом кластеризации, а типы отношений между понятиями – объектами распознавания. Тематическая структура представляет собой новый информационный объект, позволяющий: структурировать множество документов; осуществлять навигацию по этому множеству; анализировать информацию, относящуюся к структурным компонентам предметной области в соответствии с ее значимостью; повысить уровень обобщения информации.

Ключевые слова: интеллектуализация информационно-поисковых систем, тематическая структура предметной области, кластерный анализ, распознавание образов, термины индексирования, семантические отношения, новый информационный объект.

CREATION OF SUBJECT DOMAIN'S THEMATIC STRUCTURES

Vasina E.N. [1], **Kozlova I.V.** [1]

[1]*Plekhanov Russian University of Economics, 117997 Russian Federation, Moscow, Stremjannyj per., 36, e-mail: vasina_e@list.ru, ivkozlova10@mail.ru*

The toolsreview of search results representation as the form of subject domain's classification schemes or various thematic structures is provided. It is shown that tendencies of development of search engines consist in gradual expansion of traditional functions due to intellectual analytical opportunitiesconnectionto search mechanisms. The formal and mathematical problem definition of the subject domain'sthematic structure explication from the set of documents received as an information searchresult is considered. The model and process of thematic structure creation due to a terminological networkanalysis and establishment of the semantic relations between concepts are described. The terminological network is the object of a clustering, and types of concepts relations – objects of recognition.The thematic structure represents the new information object allowing: to structure a set of documents; to carry out navigation on this set; to analyze information relating to structural components of subject domain according to its importance; to raise level of information synthesis.

Keywords: intellectualization of information retrieval systems, subject domain's thematic structure, cluster analysis, pattern recognition, indexing terms, semantic relations,new information object.

Принципиально важным фактором, определяющим направление развития современных информационных систем, является неуклонное возрастание объемов информации. Даже отфильтрованные информационными системами результаты поиска представляют собой огромные потоки документальной информации. Это заставляет создателей автоматизированных информационно-поисковых систем (АИПС) все больше внимания уделять развитию инструментария представления результатов поиска в виде классификационных схем предметных областей или различных тематических структур. Приведем несколько примеров.

Интеллектуальная поисковая система Nigma [12] осуществляет автоматическую(на

основе семантического анализа) кластеризацию результатов поиска, выдаваемых другими поисковыми системами Интернет (Google, Yahoo, MSN, Yandex, Rambler). Найденные документы разбиваются на кластеры, представленные в виде иерархического дерева. С помощью установки фильтров можно отсечь ненужные темы, что облегчает поиск необходимой информации. Просматривая описания кластеров, пользователь выбирает наиболее интересные для углубленного изучения.

Vivisimo (clusty) [8] – еще одна метапоисковая система, использующая другие поисковые системы для предварительного отбора текстовой информации споследующей кластеризацией результатов поиска. Алгоритмы работы vivisimo основаны на использовании стандартной модели работы с ключевыми словами и кластеризации результатов поиска. Группировка предварительно отобранных документов проводится по трем критериям:

- по частоте встречаемости ключевых слов в результатах поиска;
- по поисковым системам, в которых были найдены документы;
- по доменным зонам (например, com, ru и др.).

Результаты кластеризации по ключевым словам представляются в виде списка пунктов меню, по которым возможна пересортировка результатов выдачи. При отображении документы кластера упорядочиваются по статистике найденных в них ключевых слов.

Аналогичный принцип отображения результатов кластеризации реализован австралийским поисковым сервером Mooter [11], на котором применяется визуальный подход к предоставлению результатов поиска по обрабатываемым запросам путем группировки результатов первичного поиска по категориям.

Другой поисковый сервер iBoogie [9] также группирует результаты поиска, но отображает их в виде, близком к экрану Проводника Windows.

Система контент-мониторинга InfoStream [10] применяется для решения задач автоматизированного сбора информации с открытых web-сайтов, ее обработки, систематизации и обеспечения доступа к ней в поисковых режимах. Одним из преимуществ системы по сравнению с традиционными сетевыми информационно-поисковыми системами является наличие аналитического инструментария, который позволяет пользователю в режиме реального времени не только получать результаты поиска, но и формировать дайджесты, строить сюжетные цепочки, анализировать взаимосвязь рубрик, динамику понятий и т.д.

Независимо от формы представления результатов, поисковые системы Интернета выдают список ссылок на найденные страницы. Пользователь при этом вынужден заниматься навигацией по найденным ссылкам, анализом страниц и поиском необходимой информации. Семантические поисковые системы AskNet [7] обеспечивают вывод ответов на

запросы пользователей непосредственно на страницу результатов поиска.

В справочно-информационной системе ВИНИТИ [13] вывод результатов поиска осуществляется поэтапно. После проведения поиска формируется сообщение, содержащее текст запроса, дату поиска, имя БД, в которой проводился поиск, сведения о количестве найденных документов и гиперссылку для перехода на просмотр краткой формы описания документов. Это сообщение записывается в историю поиска, которая отражается на экране. После анализа результатов поиска в краткой форме и выбора условий вывода на экран выводится выбранная форма документов.

В [1] описывается методика автоматической рубрикации, которая используется для распределения результатов поиска по определенным темам в поисково-аналитической системе «Галактика-Зум». Предварительно системой определяются информационные портреты (ключевые темы конкретных рубрик) по оригинальной технологии выделения и ранжирования ключевых тем. Затем автоматически происходит классификация документов методом сравнения информационных портретов документа и заданных рубрик.

Таким образом, тенденции развития поисковых систем заключаются в постепенном расширении традиционных функций и активном подключении к поисковым механизмам интеллектуальных аналитических возможностей. Один из способов интеллектуализации АИПС состоит в представлении результатов поиска в виде тематических структур (ТС) предметных областей, в качестве которых рассматриваются области научных исследований.

Задача построения (ТС) предметных областей основывается на:

-формализованном представлении тематической структуры как упорядоченной совокупности понятий предметной области и отношений между ними;

-оценке совместной встречаемости терминов индексирования в документальных БД;

-анализе и обобщении семантических элементов.

В основе построения тематической структуры лежат следующие принципы:

1. Модель тематической структуры области исследований представляется в виде кортежа множеств:

$$\mathbf{\Omega = <P, V, R>,}$$

где P – множество понятий предметной области; V – множество свойств понятий; R – множество отношений из PxV.

Используются идеи аксиоматической теории сходства, устанавливаются критерии сходства понятий в локальном и глобальном смысле [5].

2. Формализация представления тематического сходства понятий в рамках направления области исследований основана на использовании глобального сходства, определяемого общностью свойств на подмножестве понятий.

3. Для структуризации понятий тематических направлений используются семантические отношения иерархического (род – вид, целое – часть, проблема – аспект) и неиерархического типа (объект – метод, объект – область применения и т. д.).

Предполагается, что тематическая структура имплицитно содержится в выборке документов, полученной в результате поисков по запросам пользователя, являющейся моделью предметной области исследований:

B=<T, D, R'>- модель предметной области,

где T – множество терминов индексирования; D – множество документов; R' – множество отношений изTxD.

В этом случае задача экспликации состоит в поиске способа отображения модели тематической выборки документов в модель тематической структуры области исследований **Ω**, т. е. **w: B → Ω**. Исходя из этого, можно наметить два этапа решения этой задачи:

- на первом этапе из заданного множества понятий **P** необходимо выделить группы тематически связанных понятий. В формальной постановке это соответствует задаче классификации объектов-понятий и требует задания сходства между понятиями, а также выбора метода группирования;
- на втором этапе решения задачи проводится упорядочение понятий внутри выделенных групп и придание им определенной структуры в соответствии с заданным типом отношений.

В основе формальных методов классификации лежит отношение сходства между классифицируемыми объектами, при этом пользуются попарным сравнением объектов, т.е. отношение сходства рассматривается как бинарное. Аксиоматическая теория сходства рассматривает понятие сходства как отношение толерантности – рефлексивное и симметричное бинарное отношение. Для структуризации тематической области используется формализованное представление локального сходства между терминами индексирования в документах выборки.

Рассмотрим множество $T=\{ t_1, t_2..., t_i,..., t_N\}$ терминов индексирования множества документов $D = \{d_1, d_2,..., d_j,..., d_M \}$. На множестве T будем считать заданным набор признаков (свойств), т.е. одноместных предикатов вида $P(t_i)$, принимающих значения 0 или 1. Если $P(t_i) = 1$, то будем говорить, что t_i обладает признаком P_i. В качестве множества всех рассматриваемых признаков в данном случае принимается множество документов D. Тогда соответствие **f: T→D** устанавливает для каждого t_i все признаки, которыми обладает термин t_i (все документы, заиндексированные термином t_i). Это множество признаков будем обозначать $D(t_i)$, $D(t_i) \subseteq D$. Обратное соответствие -$\mathbf{f^{-1}: D \rightarrow T}$- сопоставляет каждому признаку

d_j множество $T(d_j)$ тех терминов, для которых выполнен этот признак. Соответствие устанавливает отношение на множествах терминов T и документов D и определяется как подмножество R декартова произведения множеств TxD, $R \subseteq TxD$.

Рассмотрим тройку <T, D, R>, где T – множество объектов (терминов), D – множество признаков (документов), R – отношение из TxD. Будем называть упорядоченную тройку C=C<T, D, R> картой [13]. Таким образом, карта – это экспликация понятия «множество с признаками». Вхождение множества T в карту (т. е, задание на T признаков) позволяет определить на множестве T отношение локального сходства. Отношение τ на множестве T является отношением толерантности (сходства) при соблюдении следующих условий:

- $t_i \tau\ t_i . t_i \in T$ – рефлективность;
- $t_i\ \tau\ t_j . \rightarrow t_j\ \tau\ t_i\ , t_j \in T, t_i \in T,$ – симметричность;
- $t_i \tau t_j . \& t_i \tau t_k .$ не $\rightarrow t_i\ \tau\ t_k$ – интранзитивность.

Объекты (термины) локально сходны, тогда и только тогда, когда:

$$\mathbf{D(t_i) \cap D(t_j) = \emptyset,}$$

т.е. локальное сходство требует наличия общего признака у пары терминов (документов) и является бинарным однородным отношением.

Множество T с заданным на нем отношением сходства τ является пространством толерантности $\mathbf{T}^{\tau} = \mathbf{< T, \tau >}$. Его можно изобразить неориентированным графом $G(T, \tau)$ – терминологической сетью, в которой ребрами соединены только те вершины, которые связаны отношением τ. Преобразование локального сходства пары терминов в глобальное сходство подмножества терминов может быть достигнуто двумя путями: либо построением классов толерантности на множестве T, либо установлением отношения транзитивного замыкания τ на множестве T [1].

Таким образом, задача выделения на графе классов толерантности или поиска транзитивного замыкания отношения τ, рассматриваемая нами как задача разбиения терминологической сети, моделирующей тематическую область исследований на отдельные составляющие (направления) – подмножества связанных тематическим сходством терминов, сводится к разбиению графа $G(T, \tau)$ на максимально полные подграфы или связные его компоненты [4].

---

[1] Под транзитивным замыканием (или просто замыканием) отношения τ понимается бесконечное объединение $\boldsymbol{\tau^i}$. Обозначим замыкание как $\boldsymbol{\tau}$ *, тогда $\boldsymbol{\tau}^* = \boldsymbol{\tau}^1 \cup \boldsymbol{\tau}^2 \cup \ldots \cup \boldsymbol{\tau}^k$

Следующим этапом решения задачи построения тематической структуры является структуризация терминов внутри выделенного направления. Для этого определяются основные виды (классы) семантических отношений и находятся статистические характеристики их появления в предметной области. Затем с помощью статистических критериев решается вопрос о принадлежности каждой пары терминов одному из заданных классов отношении.

Для этого используются методы кластерного анализа и распознавания образов [2, 3], причем терминологическая сеть G(T, τ) рассматривается как объект кластеризации, а тип отношений между терминами индексирования является объектом распознавания.

В результате анализа методов кластеризации и особенностей их использования для структуризации тематических областей выбрана односвязывающая кластер–процедура [2]. При этом методе достаточно одного звена, чтобы вся цепь оказалась собранной, что позволяет учитывать сходство терминов при формировании тематического направления не только по их совместной встречаемости, но и по сходству их окружения. Полученные связные компоненты не пересекаются, т.е. каждый термин присутствует только в одной группе, что приводит к четким границам отдельных направлений, выделяемых в тематической области.

Ситуация, возникающая при анализе пар терминов, связанных определенными семантическими отношениями и состоящая в обнаружении и выделении признаков, характеризующих эти пары, а затем в отнесении каждой пары к одному из заданных классов отношений, аналогична ситуации, возникающей в системах распознавания образов. Анализ семантических отношений между терминами индексирования в базе данных предполагает установление типа отношений для каждой пары терминов, упорядочение и структуризацию терминов на основе определенного типа отношений. Решение этой задачи становится возможным на основе выявления устойчивых отношений между терминами и статистических закономерностей их появления с увеличением объема БД.

Сформированная на основе множества документов, полученного в результате поискового процесса, тематическая структура представляет собой новый информационный объект, который позволяет:

- структурировать полученное в результате поиска множество документов;
- осуществлять навигацию по этому множеству;
- анализировать информацию, содержащуюся в полученных документах, относящихся к структурным компонентам предметной области в соответствии с их значимостью;
- решить проблему дальнейшего повышения уровня обобщения информации.

Список литературы

1. Антонов А.В., Курзинер Е.С. Определение тематически значимых документов в системе галактика-zoom (авторубрикация) [Электронный ресурс]. – М.: Корпорация «Галактика», 2010. – URL: http://www.dialog-21.ru(дата обращения: 18.10.2013).
2. Дюран Б., Оделл П. Кластерный анализ. – М.: Статистика, 1977. – 127 с.
3. Журавлев Ю.И., Рязанов В.В., Сенько О.В. Распознавание. Математические методы. Программная система. Практические применения. – М.: Фазис, 2006.
4. Оре О. Теория графов. – М.: Наука, 1980. – 336 с.
5. Финн В.К. Об интеллектуальном анализе данных // Новости Искусственного интеллекта. – 2004. - № 3.
6. Шрейдер Ю. А. Алгебра классификации // НТИ. Сер. 2. – 1974. - № 9. – С. 3-6.
7. Поисковая система AskNet [Электронный ресурс]. – URL:www.asknet.ru (дата обращения: 18.10.2013).
8. Поисковая системаClusty [Электронный ресурс]. – URL: www.clusty.com (дата обращения: 18.10.2013).
9. Поисковая системаiboogie [Электронный ресурс]. – URL:www.iboogie.com(дата обращения: 18.10.2013).
10. Поисковая системаInfoStream [Электронный ресурс]. – URL:www.infostream.ua (дата обращения: 18.10.2013).
11. Поисковый сервер Mooter [Электронный ресурс]. – URL:www.mootermedia.com (дата обращения: 18.10.2013).
12. Поисковая система Nigma [Электронный ресурс]. – URL:www.nigma.ru (дата обращения: 18.10.2013).
13. Поисковая система ВИНИТИ [Электронный ресурс]. – URL:www.viniti.ru (дата обращения: 18.10.2013).

Рецензенты:

Романов В.П., д.т.н., профессор, профессор кафедры информатики РЭУ им. Г.В. Плеханова Минобрнауки РФ, г. Москва.

Колмаков И.Б., д.э.н., к.ф.-м.н., профессор, профессор кафедры информатики РЭУ им. Г.В. Плеханова Минобрнауки РФ, г. Москва.