# ОБ ОРГАНИЗАЦИИ МОДЕЛЬНОЙ ПОДДЕРЖКИ ИНФОРМАЦИОННОГО ПОИСКА В РАМКАХ ОНТОЛОГИЧЕСКИ-ОРИЕНТИРОВАННОГО ТЕМАТИЧЕСКОГО ПОРТАЛА

**проф. Верлань А.Ф., доктор технических наук**

*Институт проблем моделирования в энергетике им. Г.Е.Пухова, Киев, Украина*

*a.f.verlan@gmail. com*

**доц. Олецкий А.В., кандидат технических наук**

*Национальный Университет «Киево-Могилянская Академия», Киев, Украина*

*oletsky@ukma. kiev. ua*

**Резюме: Рассматриваются возможности, связанные с организацией модельной поддержки информационного поиска на тематическом портале. Предлагается подход, связанный с погружением графа "онтология-документ" в сеть взаимосвязанных моделей.**

**Ключевые слова: информационный поиск, онтология, релевантность, модельная поддержка**

# MODEL SUPPORT FOR INFORMATION SEARCH AT THE ONTOLOGY-ORIENTED PORTAL

**prof. Verlan A.F., Doctor of Science**

*Institute for Simulation Problems in Power Engineering, Kyiv, Ukraine*

*a.f.verlan@gmail. com*

**assoc. prof. Oletsky O., PhD**

*National University "Kiev-Mohyla Academy", Kyiv, Ukraine*

*oletsky@ukma. kiev. ua*

**Abstract:** Organization of model support for information search at an ontology-oriented porta) is regarded. An approach related to integration of the model "ontology-document" into the network of connected models.

**Keywords:** information search, ontology, relevance, model support

## Введение

Проблема поиска информационных ресурсов, которые наиболее точно соответствуют целям и информационным потребностям пользователей, очень важна и актуальна, но она далека от полного решения. Существующие подходы [1-3 и др.] носят в основном эвристический характер и затрагивают только отдельные аспекты проблемы.

Не вызывает сомнений, что повышение качества поиска, полноты и релевантности его результатов должно быть тесно связано с максимальной ориентацией на семантику, онтологию предметной области. Такая онтологическая ориентированность имеет особое значение для тематически однородных веб-ресурсов, для которых характерны высокая информационная связность, достаточно высокая структурированность и качество информационного наполнения. К таким ресурсам можно отнести, например, тематические порталы, в том числе учебного назначения. При наличии развитой онтологической компоненты можно говорить уже не просто про поиск по классическим схемам, а про автоматизированный экспертный подбор наиболее адекватных информационных ресурсов, и такие системы должны приобретать экспертно-консультационный характер [4].

В работах [4, 5, 6 и др.] развивается подход на основе построения и анализа формальной модели информационного наполнения тематического веб-портала в виде графа "онтология-документ". В рамках этой модели вводится определенная система параметризованных математических соотношений, описывающих взаимосвязи между различными тематическими узлами и соответствующими им документами. В то же время остается открытым вопрос о выборе параметров этих соотношений. Очевидно, что априорный подбор этих параметров не представляется возможным, и они должны выбираться и корректироваться уже в рабочем режиме, экспериментальным путем. Возникает необходимость в создании инструментальных средств, обеспечивающих проведение подобного эксперимента и анализ его результатов [4].

В этом контексте как один из возможных подходов, направленных на проектирование и создание инструментальных средств, обеспечивающих качественную настройку системы информационного поиска, рассматривается концепция модельной поддержки информационного поиска [7]. Речь идет о создании некоторого банка моделей, описывающих различные аспекты поведения системы, а также связанных с этими моделями интеллектуальных программных средств. Сами по себе элементы системы модельной поддержки тоже могут быть

связаны друг с другом и, соответственно, образовывать те или иные структуры (например, в виде деревьев или графов). Настоящая работа направлена на рассмотрение возможных компонентов модельной поддержки информационного поиска и связей между ними.

## Основное содержание работы

Как базовая рассматривается модель информационного наполнения тематического портала в виде графа "онтология-документ", построенная на основе формальных моделей онтологии [4-6 и др.]. Эта модель в наиболее общих чертах описывается как тройка $M=<W^*,D,L>$, где $W$ - онтология предметной области, $W^*$- расширенная онтология, наполнение онтологии $W$ конкретными экземплярами классов (фактически - база знаний), $D$ - множество документов; $L$ - множество связей между $W^*$ и $D$. Собственно онтология описывается как тройка $<Q, R, F>$, где $Q$ - множество классов, которые соответствуют понятиям предметной области, $R$ - множество связей между ними, а $F$ - множество функций интерпретации. Соответственно, расширенная онтология описывается как тройка $<Q^*, R^*, F^*>$, где $Q^*$ - множество классов вместе с их экземплярами, $R^*$- множество связей между этими элементами, aF - множество функций интерпретации, определенных в простейшем случае на элементах из $Q^*$, $R^*$ та $Q^*$ x $R^*$ x $F^*$. Тогда элементы $D$ могут быть значениями функций из $F^*$. Такая формализация описывает граф "онтология-документ", узлы которого соответствуют понятиям предметной области и информационным ресурсам, а дуги - связям между ними, причем эти связи могут быть разных типов. Таким образом, если $w$ является элементом расширенной онтологии, $a$ $d$ - артефактом информационной системы, то функции интерпретации / и соответствующие весовые коэффициенты могут формироваться на основе этих категорий сущностей.

Далее, может быть осуществлен переход к модели "онтология-артефакт-пользователь-проект", в которой меры важностей связей зависят от характеристик и целей посетителей. Альтернативный взгляд на проблему может заключаться в построении многокомпонентной онтологической системы, отдельные компоненты которой соответствуют отдельным категориям сущностей [8].

Все это позволяет в общих чертах охарактеризовать наиболее важные компоненты системы модельной поддержки информационного поиска, в частности:

1. Базовая модель в виде графа "онтология-документ", описанная выше.

2. Параметризованные модели, описывающие собственно меры близости между узлами онтологии предметной области и документами. Для построения таких мер близости можно использовать ряд известных подходов на основе булевой и взвешенной векторно-матричной модели, теоретико-множественного анализа связанных элементов [1-3 и др.].

3. Параметрические модели, задающие комбинированные меры релевантности. Обозначим через $r^q(w,d)$, $q \in Q$, $w \in W$, $d \in D$ меру релевантности документа $d$ понятию $w$ по связи $q$. Здесь $W$ - множество понятий предметной области, $D$ - множество артефактов информационной системы, $Q$ - заданное множество возможных типов связей. Тогда естественно рассматривать некоторую комбинированную меру релевантности документа $d$ понятию $w$, усредненную по всем связям с учетом их весовых коэффициентов [4]:

$$R(w,d) = \sum_{q \in Q} \alpha_q r_q(w,d) \qquad (1)$$

где $a^q$ - вес (содержательно - мера важности) q-го типа связей.

4. Модели, определяющие собственно процесс принятия решений относительно динамического формирования навигационного графа, задающего возможные переходы между узлами веб-сайта.

5. Модели, характеризующие посетителей веб-сайта и позволяющие построить профили и выделить наиболее типичные группы пользователей.

6. Модели, характеризующие возможные цели посетителей. В частности, в [1] описаны такие базовые типы запросов, как информационные, навигационные и транзакционные. Эти модели позволяют относить запросы к той или иной группе и на этой основе принимать более обоснованные решения о выборе материалов, наиболее релевантных запросу.

7. Модели, описывающие поведение пользователей и их навигацию по сайту. Этот вопрос в общих чертах обсуждался в [6]. В частности, в [6] утверждается, что рассмотрение и анализ некоторого порождающего навигационного процесса позволяет получать семейство соотношений, аналогичных соотношениям классического алгоритма PageRank, но уже с учетом семантических связей на графе "онтология-документ".

8. Модели, описывающие взаимное влияние между узлами. В основе подобных моделей лежит интуитивное соображение о том, что мера важности узла может зависеть от мер важности связанных с ним узлов. В частности, этой основе в [5] описывается методика

динамического перераспределения мер важности узлов на основе организации волнового процесса распространения активации.

9. Модели, описывающие собственно процесс обучения и настройки системы. Для подбора коэффициентов соотношения (1), в частности, могут применяться генетические алгоритмы [9 и др.], хорошо зарекомендовавшие себя при решении многих переборных задач. Некоторые подходы к применению генетических алгоритмов к задаче оптимизации информационного поиска в общих чертах описаны в [10].

10. Следует обратить внимание на модели и методы Data Mining, то есть интеллектуального анализа данных; поиска закономерностей, которые объясняют имеющиеся данные; добычи знаний из сырой информации [11]. Для задачи информационного поиска на тематическом портале особое значение имеют методы Web Usage Mining [12], выделяемого как самостоятельное направление и связанного с анализом посещаемости веб-ресурсов и выявления закономерностей, которые объясняют поведение посетителей.

В частности, на основе методик Data Mining можно ставить вопрос о выборе оптимальных параметров соотношения (1). Действительно, можно предполагать, что пользователь выбирает ссылки, которые он считает наиболее перспективными, и тогда основой для выбора параметров (1) становится история фактически осуществленных переходов по ссылкам.

В рамках описываемого онтологически-ориентированного подхода можно рассматривать такие постановки задач Web Usage Mining:

- множество посетителей разбивается на кластеры или по своим профилям, или по истории навигации; для каждой группы определяются наиболее приоритетные типы связей между узлами графа "онтология-документ", и на этой основе расставляются персональные весовые коэффициенты, которые зависят от характеристик посетителей;

- на основе анализа истории переходов между узлами графа "онтология-документ" оценивается вероятность того, что находясь в узле $q$ с определенным значением характеристики $a$, посетитель перейдет по ссылке, которая соответствует типу связей $r$;

- оптимизация структуры навигационного графа с целью сокращения последовательности переходов, которые должен осуществить пользователь, чтоб достичь цели;

- эффективный подбор контекстной рекламы, связанной с ресурсами с наивысшей оценкой релевантности - то есть с теми, которые могли бы с наивысшей вероятностью заинтересовать посетителя, который в данный момент находится в заданном узле графа "онтология-документ".

Построение системы модельной поддержки информационного поиска, кроме собственно набора моделей разных типов и связанных с ними процедур, должно предусматривать организацию связей между ними. В частности, следует предусмотреть:

- объединение отдельных моделей в сеть, на основании которой можно осуществлять целенаправленные переходы между ними с целью поиска наиболее подходящих из них;

- механизмы автоматического запуска тех или иных программ, связанных с моделями как узлами модельной сети.

## Выводы

В работе в общих чертах описываются возможности, связанные с организацией модельной поддержки онтологически-ориентированного поиска на тематическом портале с целью повышения его эффективности, точности и релевантности. Базовая модель информационного наполнения портала на основе графа "онтология-документ" должна быть погружена в сеть моделей. Уточнения и дальнейшие формализации рассматриваемого подхода являются предметом дальнейших исследований.

## Литература

1. Маннинг К.Д., Рагхаван П., Шютце Х. *Введение в информационный поиск.* - М.: ООО «И.Д. Вильяме», 2011. — 528 с.

2. Гаврилова Т.А., Хорошевский В.Ф. *Базы знаний интеллектуальных систем.* - СПб: Питер, 2000. - 384 с.

3. Ландэ Д.В. *Поиск знаний в Интернет.* — М.: Изд. дом «Вильяме», 2005. - 272 с.

4. Олецький О.В. Організація онтологічно-орієнтованих засобів автоматизованого добору інформаційних ресурсів на тематичному порталі. *І/Наукові записки НаУКМА. Т.99. Комп'ютерні науки.* - К., 2009. - С. 66-69.

5. Олецький О.В. Онтологічно-орієнтований інформаційний пошук на основі хвильового процесу поширення активації. *//Наукові записки НаУКМА. Т.86. Комп 'ютерні науки.* - К., 2008. — С.50-52.

6. Олецький О.В. До проблеми моделювання потоку відвідувань на онтологічно-орієнтованому тематичному порталі. *//Моделювання та інформаційні технології. Збірник наукових праць. Спеціальний випуск.* Т.2.-К..2010. -С.321-326.

7. Верлань А. Ф., Дячук А. А., Сагатов М. В. *Модельная поддержка автоматизированного проектирования сложных технических систем.* - М., 2008.

8. Олецький О.В. Побудова багатокомпонентної онтологічної системи для автоматизованого експертного підбору навчальних матеріалів. // *Теоретичні та прикладні аспекти побудови програмних систем.* Матеріали міжнародної конференції TAAPSD'2009. Київ, 8-10 грудня 2009 р. - С.83-88.

9. Рутковская Д., Пилиньский М., Рутковский Л. *Нейронные сети, генетические алгоритмы и нечеткая логика.* - М.: Горячая линия - Телеком, 2004. - 452 с.

10. Олецький О.В. Принципи застосування генетичних алгоритмів до задачі онтологічного інформаційного пошуку. *//Наукові записки НаУКМА. Т.112. Комп 'ютерні науки.* - К., 2010. - С.49-54.

11. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. *Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP.* - СПб: БХВ-Петербург, 2007. - 384 с.

12. Гончаров М. *Web Mining - добыча знаний из World Wide Web.* http://www.spellabs.ra.