Использование ресурсов Интернета для построения таксономии

Екатерина.Черняк, Борис Миркин

Отделение Прикладной Математики и Информатики
Национальный Исследовательский Университет – Высшая Школа Экономики

Аннотация. В работе предложен двухшаговый подход к построению предметных таксономий на русском языке. На первом шаге строятся высокие уровни таксономии на основе паспортов специальностей ВАК. На втором шаге таксономические темы последовательно достраиваются новыми темами, извлеченными и отфильтрованными из дерева категорий и статей русского сегмента Википедии. Во всех расчетах используется мера сходства между строкой и текстом, основанная на аппарате аннотированных суффиксных деревьев.

Ключевые слова: Достраивание таксономий, мера сходства строки тексту, Википедия, суффиксное дерево

Введение

Таксономии, или иерархические онтологии, – это популярный инструмент для представления, хранения и использования знаний о какой-либо предметной области [1,2]. Таксономия представляет собой корневое дерево, организованное как иерархия понятий, или тем, узкой предметной области. В таком дереве темы находятся в отношении «А – часть В» или «А – более общее понятие, чем В». Автоматическое построение таксономий – важная задача, которая относится и к области автоматиче-

ской обработки текстов, и к области информационного поиска [3,4]. Наиболее популярные подходы к решению этой задачи предполагают использование больших коллекций неструктурированных текстов, относящихся к рассматриваемой предметной области. Из этих текстов извлекают ключевые слова и словосочетания, находящиеся в четко выраженном отношении «наследования», то есть, образующие искомую иерархию понятий. Недостатки такого подхода к построению таксономий хорошо известны: 1) не для каждой предметной области можно найти достаточно большую коллекцию неструктурированных текстов, 2) методы обнаружения семантических отношений между словами далеки от совершенства, поэтому охваченные темы и структура таксономии, как правило, неудовлетворительны [5]. Поэтому, в качестве замены коллекциям текстов предлагают использовать Интернет-ресурсы, например, Википедию [6]. Более того, Википедия устроена таким образом, что уже имеет некоторые задатки таксономии, например, иерархию тем (дерево категорий) и огромное количество понятий (названия статей и категорий). Однако, оказывается, что дерево категорий Википедии нельзя использовать само по себе как таксономию, поскольку качество некоторых статей или фрагментов дерева категорий вызывает сомнение в силу не профессиональности части авторов Википедии.

В данной статье представлен полуавтоматический метод построения таксономии предметной области. Он состоит из двух этапов. На первом этапе строят основу таксономии, ее два или три верхних уровня, в соответствии с официальными документами, формализующими рассматриваемую предметную области. На втором этапе происходит пошаговое достраивание тем таксономии фрагментами дерева категорий и статей из русского сегмента Википедии, очищенных от лишних тем, шума. Во всех расчетах используется мера сходства темы тексту, основанная на аннотированных суффиксных деревьях. Метод достраивания таксономии состоит из нескольких шагов. Для каждой из тем верхнего уровня: 1) находим соответствующее теме поддерево дерева категорий; 2) определяем релевантность теме статей Википедии, принадлежащих к выбранному поддереву; 3) проводим очистку поддерева от иррелевантных статей и категорий; 4) извлекаем ключевые слова и словосочетания из оставшихся после очистки статей; 5) достраиваем тему таксономии оставшимися категориями и статьями; 6) помещаем извлеченные слова и словосочетания на последний уровень таксономии в качестве уточняющих описаний листовых тем. Таксономия, построенная таким образом, отличается от большинства таксономий тем, что она сбалансирована как и по глубине, так и по числу детей и листьев в разных разделах таксономии, поскольку эти параметры контролируются в ходе построения. Ниже приведено подробное описание метода и его применение к

математической предметной области «Теория вероятностей и математическая статистика», иллюстрирующее и достоинства, и недостатки метода.

Потребность в построении таксономий математических областей вызвана нашей предыдущей работой по использованию таксономий для визуализации и интерпретации аннотаций математических статей и учебных программ по математики и информатике. Единственная доступная таксономия математики на русском языке – это рубрикатор РЖ «Математика», модифицированный последний раз в 1999 г.. Стоит заметить, что таксономия не только устарела, но и нелогична и не сбалансирована. Так, например, в ней отсутствует одна из основных тем – «Дискретная математика», под темой «Дифференциальные уравнения» находятся более 80 других тем, а под более современной темой «Теория игр» – всего 10.

К счастью, существуют и другие русскоязычные классификации научных тем, например, классификации ВАКа представляющие собой двухуровневые деревья, покрывающие все научные темы. Еще несколько уровней достаточно общей классификации можно извлечь из паспортов специальностей ВАК. Однако, для того, чтобы дойти до таких частных понятий, как «производная» или «функция», требуется еще от двух до четырех уровней менее общих понятий.

Отсюда вытекает постановка рассматриваемой задачи. Нам требуется метод достраивания таксономии на основе ресурсов Википедии. На выходе этого метода должно получаться более или менее сбалансированное дерево, глубина и число детей на разных уровнях в разных разделах не сильно бы различалось. Дополнительное требование к таксономии заключается в наличии уточнений, то есть, множества слов или словосочетаний, объясняющих листовые темы.

Требованиям сбалансированности и наличия уточнений удовлетворяет классификационная система Ассоциации Вычислительной техники ACM-CCS, которую по праву можно считать золотым стандартом таксономии.

Методам достраивания таксономии посвящено несколько работ. В исследовании [8] используются результаты поиска шаблонных запросов вида “A состоит из ...” . В результате поиска такого словосочетания предполагается получить множество понятий, которые можно рассматривать как потенциальные подтемы понятия A. Можно использовать и уже существующие онтологии и таксономии для достраивания исходной. Если и достраиваемая таксономия, и используемая в качестве источника онтология или таксономия описаны формальными моделями языка OWL, эта задача решается с помощью агрегации и введения новых логических отношений, как это сделано в работе [9]. В целом, для

достраивания таксономии могут быть выбраны как структурированные, так и неструктурированные источники. Во многих исследованиях [10-12] используется компромиссное решение: данные для достраивания таксономии извлекают из полуструктурированной интернет-энциклопедии Википедии. Многообразие данных в Википедии (инфобоксы, тексты статей, дерево категорий) и универсальность тематики энциклопедии позволяют использовать ее для построения таксономий и онтологий разных предметных областей. В работе [5] приведены и другие аргументы в пользу Википедии как источника данных для построения таксономии:

• Википедия постоянно обновляется, поэтому таксономию легко обновлять и
пополнять;

• Википедия многоязычна, и, скорее всего, методика построения таксономий, разработанная для одного языка подойдёт и для другого.

В работах [10-12] представлены различные способы построения или пополнения таксономий на основе ресурсов Википедии. В [10] в качестве источника таксономических тем использованы тексты статей, в [11] – дерево категорий Википедии, а в [12] – инфобоксы. Основное отличие нашего подхода заключается в использовании и структурированных данных – фрагментов дерева категорий Википедии, и неструктурированных текстов статей, что позволяет нам следовать золотому стандарту ACM-CCS. Мы специально решили ограничиться такой узкой предметной областью, как теория вероятностей и математическая статистика, чтобы, во-первых, избавиться от проблем обработки больших объемов данных из Википедии, во-вторых, получить на выходе таксономию разумного размера и иметь возможность скорректировать ее вручную. Кроме того, использование структурированных данных, таких как дерево категорий Википедии, в качестве источника тем для достраивания, позволяет избежать трудностей с поиском шаблонов и низкой точностью результатов такого поиска.

Метод достраивания таксономии на основе ресурсов Википедии

Мы задали основу таксономии, основываясь на паспортах специальностей ВАК [14]. В этих паспортах содержатся двух- или трехуровневые деревья тем специфических областей математики, в том числе, теории вероятностей и математической статистики. Извлеченное из соответствующего паспорта дерево (Таблица 1) и представляет собой небольшое трехуровневое дерево, на первом уровне которого расположе-

ны 2 раздела: теория вероятностей, математическая статистика. В первом разделе 5 листьев, во втором 6.

ТВиМС.01	Теория вероятностей	
	ТВиМС. 01.01	Модели и характеристики случайных явлений
	ТВиМС. 01.02	Распределения вероятностей и предельные теоремы
	ТВиМС. 01.03	Комбинаторные и геометрические вероятностные задачи
	ТВиМС. 01.04	Случайные процессы и поля
	ТВиМС. 01.05	Оптимизационные и алгоритмические вероятностные задачи
ТВиМС.02	Математическая статистика	
	ТВиМС. 02.01	Методы статистического анализа и вывода
	ТВиМС. 02.02	Статистические параметры и их оценивание по выборке
	ТВиМС. 02.03	Статистические критерии и проверка статистических гипотез
	ТВиМС. 02.04	Временные ряды и случайные процессы
	ТВиМС. 02.05	Машинное обучение
	ТВиМС. 02.06	Многомерная статистика и анализ данных

Табл 1. Основа таксономии теории вероятностей и математической статистики

Мы использовали соответствующую категорию Википедии, то есть, “Теория вероятностей и математическая статистика”, в качестве единственного источника тем для достраивания. Заметим, что мы не обращались к другим разделам Википедии, поскольку название и тематика

данной категории полностью совпадает с названием и тематикой достраиваемой таксономии. По данным на конец 2011 года категория “Теория вероятностей и математическая статистика” насчитывала 48 подкатегорий и 640 статей.

Мы пользовались двумя типа данных из доступных данных Википедии:

1. Иерархической структурой дерева категорий Википедией
2. Коллекцией неструктурированных текстов статей Википедии.

Дерево категорий мы использовали для наращивания дерева исходной таксономии: к каждой таксономической теме на первом и втором уровне присоединим несколько подходящих по содержанию категорий Википедии и лежащие ниже подкатегории. Листьями таксономии считали названия статей, а извлеченные из текстов статей ключевые слова – уточнениями листьев.

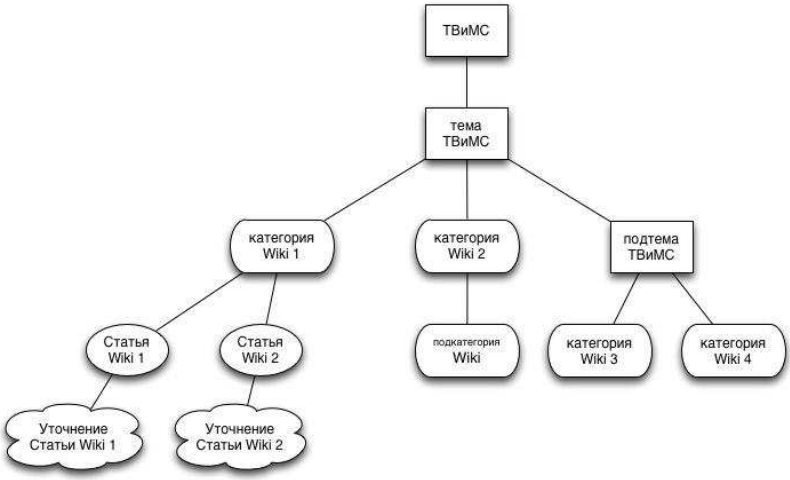

Рис. 1. Схема достраивания таксономии. Прямоугольники – темы исходной таксономии, скругленные прямоугольники – категории и подкатегории Википедии, овалы – статьи Википедии, облака – уточнения листьев, состоящие из ключевых слов и словосочетаний.

Таким образом, что каждую тему таксономии следует достроить двухуровневым деревом, состоящим из категории Википедии и принадлежащих к ней статей.

Структура категорий Википедии оказалось зашумленной: между некоторыми подкатегориями и категориями может не быть смысловой связи (например, категория Оптимизация находится в категории Машинное

обучения, которая, в свою очередь находится в категории Математическая статистика). Строго говоря, дерево категорий является решеткой, а не деревом, поскольку в некоторых случаях может содержать циклы. Одно из объяснений этому феномену согласно [15] заключается в том, что стандарт разметки Википедии допускает размещение подкатегории или статьи в неограниченном числе категорий, а авторы Википедии склонны, как правило, помещать статьи и подкатегории в как можно большее число категорий. Таким образом, для достраивания таксономии необходимо провести предварительную обработку данных из Википедии и очистить дерево категорий от иррелевантных статей и категорий.

Основные этапы достраивания таксономии:

1. Определить таксономическую тему для достраивания
2. Извлечь из Википедии фрагмент дерева категорий и статьи, соответствующие достраиваемой теме
3. Очистить дерево категорий от иррелевантных статей
4. Очистить дерево категорий от иррелевантных подкатегорий
5. Достроить следующий уровень таксономии под выбранной таксономической темой релевантными категориями
6. Достроить каждую категорию релевантными статьями – новыми листьями в таксономии
7. Извлечь ключевые слова и словосочетания из статей и использовать их в качестве уточнений листьев.

Опишем каждый из этапов на примере таксономии теории вероятностей и математической статистики (ТВиМС) и соответствующей категории Википедии.

1. Определение таксономической темы для достраивания: теория вероятностей и математическая статистика.

2. Извлечение из Википедии фрагмент дерева категорий и статьи, соответствующие достраиваемой теме. В Википедии существует категория с таким же названием, т.е. этот этап совершается путем загрузки поддерева категорий с корнем в категории “Теория вероятностей и математическая статистика”. В результате в нашем распоряжении оказались 640 статей Википедии, организованных в 48 категорий. Максимальная глубина загруженного дерева – 5, средняя глубина – 3. Некоторые категории содержат только подкатегории, но, в большинстве случаев, категории содержат и статьи, и подкатегории.

3. Очистка дерева категорий от иррелевантных статей. Среди вершин этого дерева оказались некоторые, очевидным образом не связанные с теорией вероятностей и математической статистикой, например, “Оптимизация программного обеспечения” или “Natural Language Toolkit”. Чтобы определить, релевантна ли статья категории, в которой она нахо-

дится, мы рассчитывали степень сходства названия категории с текстом статьи и, если найденная степень сходства ниже определенного порога, считали статью иррелевантной. Мы оценивали степень сходства с помощью метода аннотированного суффиксного дерева, описанного ниже. Согласно методу АСД, мера сходства, изменяющаяся от 0 до 1, выражает среднюю условную вероятность появления символа в строке после префикса строки. Чем ниже мера сходства, тем меньше вероятность, что название категории связано с содержанием статьи. Так, например, 7 из 12 статей в категории "Факторный анализ" оказались иррелевантными согласно сформулированному принципу, среди них "Линейная регрессия на корреляции" и "RANSAC".

4. Очистка дерева категорий от иррелевантных категорий. Связь между данной категорией и ее подкатегорией определяется по аналогии со связью между категорией и статьей в ней. Вместо текста статьи мы использовали совокупность всех текстов в подкатегории и рассматривали их как один текст. Степень сходства названия категории с таким текстом должна превышать заданный порог, чтобы подкатегория была релевантной для своей родительской категории. К сожалению, такой принцип определения связи между подкатегориями и категориями не всегда эффективен. Так, подкатегория "Деревья принятия решений" оказалась иррелевантной категории "Машинное обучение", поскольку ни одна из четырех статей в ней не содержит ни строки "Машинное обучение", ни ее подстрок.

5. Достраивание следующего уровня таксономии под выбранной таксономической темой релевантными категориями. После очистки дерева категорий от иррелевантных статей и категорий мы достраивали оставшиеся категории под соответствующие таксономические темы существующей таксономии. Для определения соответствия категорий таксономическим темам снова был использован метод АСД. Мы вычисляли степень сходства между таксономическими темами и текстами статей в каждой категории, слитых в один текст. Поскольку топология дерева категорий не учитывалась, в некоторых случаях к одной и той же таксономической теме были достроены и категория, и ее подкатегории. Например, категории "Непрерывные распределения" и "Дискретные распределения", принадлежащие к категории "Распределения вероятностей", были достроены к таксономической теме "Распределения вероятностей и предельные теоремы" вместе со своей родительской категорией. В этом случае мы создали новый уровень в таксономическом дереве под темой "Распределения вероятностей" и поместили на него обе подкатегории, сохранив при этом структуру дерева категорий.

6. Достраивание категорий релевантными статьями. В достраиваемой таксономии листьями служат названия статей Википедии. Пусть к так-

сономической теме достроена некоторая категория Википедии. Все релевантные статьи, оставшиеся в этой категории после очистки, помещаются на последний уровень таксономического дерева и их названия становятся листьями таксономии.

7. Извлечение ключевых слов и словосочетаний из статей и использование их в качестве уточнений листьев. Каждый лист таксономии следует снабдить множеством слов и словосочетаний, описывающих его содержание так, как это сделано в таксономии ACM-CCS. Мы использовали в качестве уточнений ключевые слова и словосочетания, извлеченные из текстов статей. Для их извлечения мы не пользовались сложными алгоритмами и считали, что ключевое слово – это существительное, частота которого в тексте статьи достаточно велика, а ключевое словосочетание – это частотная пара слов, удовлетворяющая синтаксическим шаблонам “прилагательное + существительное” или “существительное + существительное”. Таким образом, лист “Корреляция” получил такое уточнение: “коэффициент корреляции”, “случайная величина”, “ранг” и т.д., а лист “Метод максимального правдоподобия” – “функция правдоподобия”, “параметр”, “выборка”.

Метод АСД

Суффиксное дерево – это структура данных, используемая для хранения и поиска символьных строк и их фрагментов [16]. В некотором смысле, суффиксное дерево можно считать альтернативой векторной модели представления текстов (VSM) [17]. Если текст представлен суффиксными деревом, то его элементами оказываются не отдельные слова (как это происходит в наипростейшей модели “мешок слов”), а строки неограниченной длины, которые могут быть как и фрагментом слова, так и целым словом, словосочетанием и даже предложением.

Аннотированое суффиксное дерево (АСД) – это суффиксное дерево, узлы (а не ребра!) которого аннотированы частотами фрагментов строк. Алгоритм построения АСД и использования его в задаче фильтрации спама описан в [18], а в [6, 19] представлены другие приложения метода АСД.

В наших расчетах любая статья Википедии считалась множеством строк, состоящих из трех слов, а название статьи включалось в это множество без изменений. Для того, чтобы оценить сходство отдельной строки со множеством строк, мы, во-первых, строили АСД для множества строк, во-вторых, находили все совпадающие фрагменты данной строки в построенном АСД. После этого мы вычисляли оценку каждого совпавшего фрагмента: среднюю частоту символа в фрагменте, нормированную длинной всего фрагмента. Общая оценка строки вычислялась

как усредненная оценка всех совпавших фрагментов. Таким образом, окончательная оценка лежит между 0 и 1 и может считаться условной вероятностью. По сравнению с мерами сходства, описанными в [18] данная оценка имеет естественную вероятностную интерпретацию и независима от длины оцениваемой строки.

Результаты

В результате уточнения таксономии, представленной в таблице 1, была получена таксономия, глубина которой изменяется от 4 до 7. Фрагмент достроенной таксономии представлен на Рисунке 2. На этапе очистки из дерева категорий Википедии было удалено 100 иррелевантных статей и 2 иррелевантные категории. Некоторые таксономические темы, например, “Методы статистического анализа и вывода” не были достроены.

Основной недостаток построенного таксономического дерева – это положение темы “Деревья принятия решением”. Согласно представленному методу, это тема должна быть потомком темы “Многомерная статистика и анализ данных”, то есть, иметь общего родителя с темой “Машинное обучение”. Объяснение этому приведено выше: мера сходства строки “Машинное обучение” со статьями в категории “Деревья принятия решений” удивительно мала.

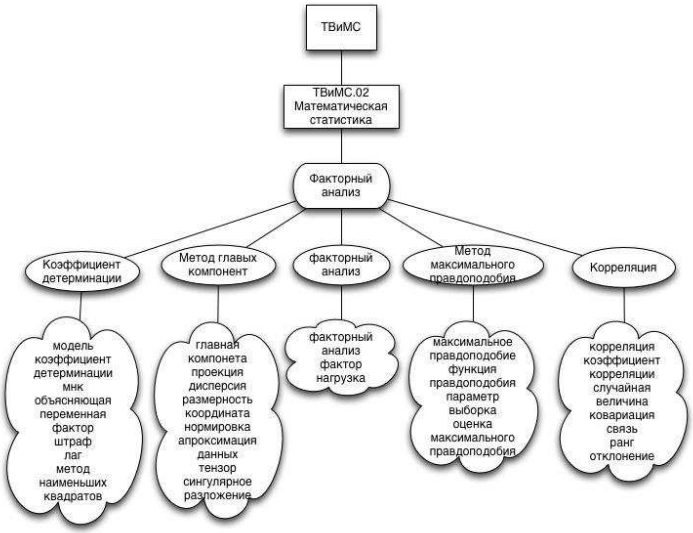

Рис. 2. Фрагмент достроенной таксономии: “Факторный анализ”

В задаче достраивания таксономии метод АСД использовался трижды:

1. Для очистки дерева категорий Википедии от иррелевантных статей;
2. Для очистки дерева категорий Википедии от иррелевантных категорий;
3. Для определения связей между таксономическими темами и категориями Википедии.

В двух первых случаях требовалось задать значение порога отсечения иррелевантных статей и категорий. Эксперименты показали, что разумно установить порог на уровне 0.2 как 1/3 от максимального получаемого значения.

Заключение

Метод автоматического достраивания таксономии – часть двухшагового подхода к построению таксономий. На первом шаге эксперт задает основу таксономии. На втором шаге таксономия автоматически достраивается тема за темой до необходимого уровня детализации. Этот подход позволяет защитить таксономию от зашумления и избежать появления иррелевантных или слишком узких тем. Википедия оказалась хорошим источником тем для достраивания, поскольку содержит и структурированные (дерево категорий), и неструктурированные (тексты статей) данные.

Использование метода АСД в задаче достраивания таксономий обладает своими достоинствами и недостатками. К его достоинствам относится независимость от языка и его грамматики. Главный недостаток метода заключается в том, что метод основан на посимвольных и пословных совпадениях и не позволяет использовать синонимы. Развитие метода будет направлено на использование синонимических отношений. Кроме того, мы будем рассматривать и другие источники таксономических тем: ГОСТы, научные статьи или учебные программы, выдачу поисковых систем по запросам – таксономическим темам.

Список источников

1. ACM Computing Classification System (ACM CCS), (1998), available at: http://www.acm.org/about/class/ccs98-html.
2. Chernyak E. L., Chugunova O. N., Mirkin B.G. (2012), Annotated suffix tree method for measuring degree of string to text belongingness [Metod annotirovannogo suffiksnogo dereva dlja otsenki stepeni vhozhdenija strok v tekstovie dokument], Biznes-Informatika [Business Informatics], no.3, pp. 31-41.
3. Chernyak E.L., Chugunova O.N., Askarova J.A., Nascimento S., Mirkin B. G. Abstracting concepts from text documents by using an ontology. Proceedings of the 1st International Workshop on Concept Discovery in Unstructured Data. Moscow, 2011, pp. 21-31.
4. Grau B.C., Parsia B., Sirin E. Working with Multiple Ontologies on the Semantic Web. In Proceedings of the 3d International Semantic Web Conference, Hiroshima, Japan 2004, pp. 620-634.
5. Grineva M., Grinev M., Lizorkin D. (2009) Text documents analysis for thematically grouped key terms extraction, in Trudy Instituta sistemnogo programmirovanija RAN, Institute for System Programming, pp. 155-156 (in Russian).
6. Gusfield D. (1997), Algorithms on Strings, Trees, and Sequences, Cambridge University Press.
7. Higher Attestation Commission of RF Reference, (2009), available at: http://vak.ed.gov.ru/ru/help_desk/.
8. Kittur A., Chi E.H., Suh B. What's in Wikipedia? Mapping topics and conflict using socially annotated category structure. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Boston, USA, 2009, pp. 1509-1512.
9. Liu X., Song, Y., Liu S., Wang H. Automatic Taxonomy Construction from Keywords. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, USA, New York, 2012, pp. 1433-1441.
10. Loukachevitch N.V. (2011), Tezaurusy v zadachah informatsionnogo poiska [Thesauri in information retrieval tasks], MSU, Moscow.

11. Pampapathi R., Mirkin B., Levene M., (2006), A suffix tree approach to anti-spam email filtering, Machine Learning 2006, Vol. 65(1), pp. 309-338.
12. Ponzetto S.P., Strube M. Deriving a Large Scale Taxonomy from Wikipedia. In Proceedings of AAAI Conference on Artificial Intelligence, Vancouver, Canada, 2007, pp. 78-85.
13. Robinson P.N., Bauer, S. (2011), Introduction to Bio-Ontologies, CRC, USA.
14. Sadikov E., Madhavan J., Wang L., Halevy A.Y., Clustering query refinements by user intent. In Proceedings of the 19th International Conference on World Wide Web, New York, USA, 2008 pp. 841-850.
15. Taxonomy of Abstracting Journal "Mathematics" (1999), VINITI. Available at: http://www.viniti.ru/russian/math/files/271.htm.
16. Van Hage W.R., Katrenko S., Schreiber G., A Method to Combine Linguistic Ontology-Mapping Techniques. In Proceedings of 4th International Semantic Web Conference, 2005, Galway, Ireland, pp. 34-39.
17. White R.W., Bennett P.N., Dumais S.T. Predicting short-term interests using activity-based search contexts. In Proceedings of 19th ACM conference on Information and Knowledge Management, Toronto, Canada, 2010, pp. 1009-1018.
18. Wu F., Weld. D. Automatically refining Wikipedia Infobox Ontology. In Proceedings of the 17th International World Wide Web Conference, Beijing, China, 2008, pp. 635-645.
19. Zamir O, Etzioni. O. Web document clustering: A feasibility demonstration. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, USA, 1998, pp. 46-54.
20. Zirn C., Nastase V., Struve M., Distinguishing between Instances and Classes in the Wikipedia Taxonomy. In Proceedings of 5th European Semantic Web Conference, Tenerife, Spain, 2008, pp. 376-387.