МЕХАНИЗМЫ ОСНАЩЕНИЯ РУБРИКАТОРА ВИНИТИ КЛЮЧЕВЫМИ СЛОВАМИ

FORMAL METHODS TO SUPPLY VINITI RUBRICATOR WITH KEYWORD SETS

К.О. Малинина *(malinina@viniti.ru)*
А.В. Шапкин *(ss@viniti.ru)*
Всероссийский институт научной и технической информации РАН

Предлагаются механизмы формализации предметного описания рубрик Рубрикатора ВИНИТИ: структуризация статистических данных, нормализация терминов, обеспечение взаимодействия экспертов, развитие структуры БД. Предметное описание рубрик и кластеризация списка терминов могут быть полезны при построении поискового тезауруса по тематике ВИНИТИ.

Введение

В работе рассматриваются механизмы предметного описания рубрик Рубрикатора ВИНИТИ, т.е. оснащения их списками ключевых слов. Предметное описание рубрик предполагает трудоемкую и кропотливую ручную работу специалистов в соответствующих областях наук. Для облегчения их задачи предлагаются механизмы формализации составления списков ключевых слов, особенно в части обеспечения взаимодействия экспертов. Методика предполагает:

1) подготовку исходного материала на основе статистической обработки и анализа данных, собранных на массиве информационных продуктов ВИНИТИ с 2001 года;
2) развитие структуры базы данных (БД), предназначенной для хранения как накопленных данных, так и результатов работы экспертов с сохранением ссылки на источник появления данных;
3) структуризацию списка терминов с выделением дескрипторов т.е. нормализацию терминов;
4) разработку механизма «слияния» исходных данных и данных, полученных от разных экспертов разного уровня компетенции в данной области наук и/или тематике;
5) создание программного обеспечения для работы экспертов, а также программ анализа списков ключевых слов статистическими методами.

Предметное описание рубрик позволит:

(1) снабдить Рубрикатор ВИНИТИ пояснительной информацией, уточняющей круг тематик, связанных с рубриками;
(2) связать тематическую классификацию с предметной;
(3) создать предметный указатель к Рубрикатору ВИНИТИ;
(4) по набору ключевых слов находить соответствующие рубрики, т.е. определять тематику документа по ключевым словам.

Результаты работ по оснащению рубрик ключевыми словами и кластеризации списка терминов могут быть использованы при решении проблемы построения поискового тезауруса по естественным и техническим наукам.

Отбор рубрик, оснащаемых списками терминов

Рубрикатор ВИНИТИ представляет собой дерево с фиктивной корневой вершиной.

Уровни рубрик нумеруются сверху, от корня, вершины нулевого уровня. Первый уровень имеют рубрики, соответствующие областям наук в целом. Уровень рубрики, имеющей родительскую вершину уровня m, равен m+1.

При оснащении Рубрикатора ключевыми словами следует в первую очередь определить уровни рубрик, которые подлежат рассмотрению.

Рубрика, с одной стороны, относит документ к области наук, а с другой – определяет конкретную тематику в рамках заявленной в родительской рубрике. Рубрики нижнего уровня могут также использоваться для группировки рефератов при публикации их в Реферативном журнале (РЖ) ВИНИТИ. Поэтому нецелесообразно оснащать списками ключевых слов как рубрики, близкие к корню, так и рубрики низкого уровня. Для разных ветвей Рубрикатора могут быть выбраны разные диапазоны уровней.

Из анализа накопленных данных, а также из работ по данной теме (см. [**1**]) сделано заключение, что списками ключевых слов должны быть снабжены рубрики не выше 3–4 уровня.

Рубрики «Общие вопросы» и т.п. можно исключить из рассмотрения, а встретившиеся в них ключевые слова отнести к родительской рубрике, определяющей предметную область. Аналогично можно исключить и

рубрики «Другие вопросы» и т.п. Анализ ключевых слов, использованных в поисковых образах документов (ПОД) подобных рубрик представляет собой отдельную задачу – выявление новых направлений исследований в рамках тематики, заявленной в вышестоящих рубриках (см. [**2**]).

Массив ключевых слов

Под *ключевым словом*, или *термином*, будем понимать как отдельное слово, так и не разбиваемое на части словосочетание. Варианты написания одного и того же термина, хранящиеся в БД, объединяются в группы эквивалентности – *кластеры* словоформ.

В БД ведется список всех ключевых слов, использованных в ПОДах в РЖ, начиная с 2001 года, – *массив ключевых слов* ВИНИТИ. Для объединения ключевых слов в непересекающиеся кластеры словоформ использовались автоматизированные методы[1]. Предметизация рубрик является разбиением единого списка ключевых слов, допускающим включение одного и того же термина в разные списки.

Определить, является ли данное словосочетание результатом ошибки ввода или обработки документа, в ряде случаев можно и без привлечения специалистов. Более того, отыскание некоторого класса ошибок и их исправление может быть автоматизировано. Однако для выяснения правильности термина и его классификации с точки зрения допустимости использования в качестве ключевого слова в соответствующей области необходимо привлечение экспертов.

Ключевое слово может быть отнесено к одной из следующих категорий:

1) ошибка в написании;
2) устаревший или редко употребляемый вариант термина;
3) термин, не использующийся в данной области наук;
4) один из равноправных вариантов написания термина;
5) «канонический» вариант, наиболее часто употребляемый в данной области;
6) представление термина, предпочтительное для предметного указателя.

Поскольку удаление ключевых слов, использовавшихся в ПОДах, недопустимо, предусмотрено логическое удаление, выводящее их из выбранной области видимости.

В зависимости от способа внесения в массив, ключевые слова делятся на 3 группы:

1) загруженные «по факту использования» в ПОДах рефератов, опубликованных в РЖ;
2) внесенные пользователем при работе в клиентской программе;
3) загруженные из списка, составленного экспертом.

В массив ключевых слов ВИНИТИ также входят словосочетания, которые характеризуют не столько предметную область, сколько характер самого документ (например, «обзоры», «персоналии», «конференции»). Они либо должны быть исключены из рассмотрения вместе с рубриками «Обзоры» и т.п., либо «осесть» в этих рубриках, либо «всплыть» на самый верхний уровень.

Загрузка списка ключевых слов рубрик

Словосочетание, встречающееся в предметном описании рубрик, составленном экспертом, можно отнести к одному из 4-х видов:

1. Ключевое слово, приписанное к данной рубрике ранее. Связь между рубрикой и термином уже присутствует в БД.
2. Термин, присутствующий в массиве ключевых слов, но либо отнесенный к другой рубрике, либо не связанный ни с одной из них.
3. «Не канонический» вариант термина, уже приписанного к этой рубрике. Происходит смена представителя кластера в списке, т.к. из каждого кластера к рубрике может быть приписан только один термин.
4. Словосочетание, отсутствующее в массиве ключевых слов. Оно вносится в список со статусом «предложено экспертом». Затем осуществляется поиск других его форм. В результате оно будет либо добавлено к найденному кластеру, либо образует новый.

В случаях 3–4 может потребоваться дополнительная ручная обработка: выбор «канонического» варианта, либо разделение кластера на 2 или более новых, что вызовет необходимость анализа вхождений терминов кластера в списки других рубрик.

Наследование предметного описания рубрик

Поскольку рубрики связаны иерархически, можно говорить о наследовании их предметного описания. В зависимости от трактовки иерархических отношений возможны два подхода.

1. Рубрика может рассматриваться как объединение всех её подрубрик («взгляд сверху вниз»). Тогда предметным описанием можно считать не только список терминов данной рубрики, но и объединение ключевых слов всех её подрубрик. Такой подход может быть полезен для рубрик верхних уровней, соответствующих областям наук и их основных направлений.

[1] При объединении ключевых слов в кластеры по признаку написательной близости терминов использовались работы сотрудников ВИНИТИ Федорца О. В. [см. **4**] и Котко А. А.

2. Рубрика может рассматриваться как сужение тематики, определяемой рубрикой предыдущего уровня («взгляд снизу вверх»). В этом случае все термины, имеющие отношение к рубрике, относятся и к её подрубрикам. Следовательно, предметное описание рубрики включает в себя не только её список ключевых слов, но и объединение списков терминов для всех вышестоящих рубрик. Этот подход интересен для рубрик нижних уровней, рубрик «Обзоры», «Общие вопросы», «Другие методы» и прочих рубрик, не имеющих собственных списков ключевых слов.

Поэтому в предметное описание рубрики должны вноситься только термины, относящиеся к данному кругу вопросов, но которые не могут быть отнесены ни к более широкой, ни к более узкой тематике.

Списки терминов могут быть сокращены за счет вычищения из них ключевых слов, входящих в описания родительских рубрик. Однако допустим вариант, когда термин, входящий в описания всех подрубрик некоторой рубрики, отсутствует в её собственном предметном описании.

Классификация связи «рубрика – ключевое слово»

Классификация *источников появления* ключевого слова в предметном описании рубрики:

1. По факту использования. Пара «рубрика – ключевое слово» встретилась в документах.
2. Связь между рубрикой и термином установлена на основе автоматизированной статистической обработки.

Дополнительными параметрами в этих случаях являются частотные показатели.

3. Термин внесен в список во время сеанса работы в клиентской программе, напрямую обращающейся в БД, с проведением всех необходимых проверок.
4. Ключевое слово загружено из списка, составленного экспертом, в пакетном режиме.

Поскольку список ключевых слов рубрики может содержать несколько десятков терминов, имеет смысл дополнительно ранжировать связь по *значимости*. На начальном этапе могут быть выделены ранги:

1) основные термины, списки которых при рубриках в пределах одной верви Рубрикатора по возможности не должны пересекаться;
2) «факультативные» (редко встречающиеся) термины, которые могут быть основными для других рубрик.

При добавлении к предметному описанию рубрики списков терминов из рубрик, связанных с ней иерархически, наследуемые термины динамически помечаются как «унаследованные», их ранг может зависеть от расстояния между рубриками и от исходного ранга.

При входе в клиентскую программу пользователь выбирает область, с которой предполагает работать как специалист (рубрика 1-го уровня) и как эксперт (рубрики 2–3 уровня). Выбор тематики пользователем хранится в БД и после завершения сеанса работы программы.

При работе со списком ключевых слов выбранной рубрики, эксперт может:

- внести новый термин;
- вычеркнуть имеющийся термин;
- передвинуть ключевое слово «вверх», т.е. отнести его к родительской рубрике;
- отнести ключевое слово к рубрике другой тематики.

В последнем случае термину в списке «сторонней» рубрики будет присвоен соответствующий признак.

Специалист может быть экспертом в нескольких областях. Тогда его рекомендации имеют больший «вес», чем мнения других специалистов.

Итак, при добавлении ключевого слова в список, либо при подтверждении автоматически установленного соответствия, можно выделить 3 категории:

(1) отнесено экспертом при составлении списка ключевых слов данной рубрики или тематики;
(2) отнесено экспертом при работе со списком ключевых слов рубрики другой тематики;
(3) отнесено специалистом (экспертом в смежной области) при работе со списком ключевых слов своей рубрики или тематики.

Такое деление обосновано, т.к. мнение эксперта при работе со списком ключевых слов может отличаться от решения отнести термин к рубрике при предметизации рубрик смежной тематики.

Специалист, не являющийся экспертом, может только добавлять термины в список.

При составлении предметного описания рубрик эксперту могут понадобиться следующие сведения:

1) Источник и способ появления термина в списке. Для термина, приписанного к рубрике автоматически, полезно знать статистические характеристики и варианты его употребления, а также номера РЖ, в которых были опубликованы соответствующие рефераты. Для детального анализа желателен выборочный просмотр рефератов.
2) Рубрикатор ВИНИТИ.
3) Поиск заданного словосочетания в массиве ключевых слов и просмотр имеющихся вариантов.
4) Списки ключевых слов «соседних» рубрик.

Поиск термина

При сопоставлении словосочетаний необходимо учитывать возможность изменения порядка слов и варианты словоформ, т.е. изменения окончаний. Первым шагом является разделение словосочетания на отдельные слова.

Для поиска словосочетаний, отличающихся от заданного порядком слов, можно использовать либо перестановку слов, что удобно лишь для терминов, состоящих из 2–3 слов, либо предварительно построенный инвертированный список. При поиске по инвертированному списку приходится налагать дополнительные ограничения для того, чтобы отсечь словосочетания, содержащие слова, не входящие в искомое, а при его построении нужен список стоп-слов.

Поиск с учетом изменения словоформ осуществляется за счет приведения слов к их основам, например, с помощью усечения одним из следующих способов:

1) От каждого слова последовательно отсекается по одной букве.
2) Все слова усекаются на это заданное число символов.
3) Поддерживается список возможных окончаний, и слова усекаются на найденное в нем окончание.

Полезно установить максимально возможную длину окончания и минимальную допустимую длину основы, которая может зависеть от длины исходного слова, а также ограничение на длину найденного слова в зависимости от длины окончания.

Для случая чередующихся гласных (например, «число» – «чисел») можно дополнить операцию усечения удалением из основы последней гласной.

Для особо сложных случаев (исключений) может быть заведен словарь основ. Тогда алгоритм выделения из слова его основы в первую очередь должен обращаться к этому словарю и только в случае неудачи выделять основу, как описано выше.

При поиске в диалоговом режиме усечение слов и/или выделение основ можно предоставить пользователю.

Слова, набранные большими буквами, как правило, являются аббревиатурами, и не должны ни усекаться, ни вычищаться как стоп-слова.

При поиске необходимо учитывать множество допустимых символов, включая кодировку спецсимволов, греческих букв и букв с диакритическими знаками. Оно определяется языками терминов (русский и английский) и наличием формул (химических, математических, определяющих физические понятия).

В ключевых словах используются:

- русские и латинские буквы (строчные и прописные);
- круглые скобки, дефис, двойные кавычки, знаки препинания;
- греческие буквы (строчные и прописные) – α-частицы, Г-функция (гамма-функция) и пр.
- верхние и нижние индексы – «аннигиляция e^+e^-», «сплавы на основе Fe_3Si» и т.д.
- цифры (0–9);
- спецсимволы, кодируемые специальным образом (это математические и формульные знаки: тире, знак градуса, символы ∞, $\leq$ и др.).

В настоящее время в массиве ключевых слов используются два варианта кодировки: алфавит ВИНИТИ и LaTeX. Кроме того, при загрузке автономно составленного списка ключевых слов спецсимволы могут быть «потеряны», а греческие буквы заменены их названиями (например, «альфа» или «alpha» вместо α). Эти случаи нужно приводить к единому представлению.

Поисковый образ документа как источник данных для предметных описаний рубрик

При проведении статистического анализа необходимо помнить, что ПОД в рефератах характеризует документ, а не рубрику, к которой отнесен документ. Поэтому ключевое слово может и не относиться ни к одной из рубрик документа. По данным с 2001 года только 3% документов проиндексированы более чем 1 рубрикой. Чаще всего используются рубрики 4-го (≈41%) и 5-го (≈33%) уровней, значительно реже – рубрики уровней 6–7 или 3. И только в 1% документов встречаются рубрики более низких или более высоких уровней.

Анализ ключевых слов для некоторых рубрик (например, 411.03.02.31.23 «Астрономия; Теоретическая астрономия. Небесная механика; Общие проблемы; Теория возмущенного движения; Вращательное движение») показал, что в качестве ключевых слов могут использоваться:

1) термины предметной области, характеризующие данную рубрику, как в случае термина «движение полюсов Земли»,
2) словосочетания, уточняющие содержание документа и дополняющие классификацию, определяемую рубрикой, например, ключевое слово «математическое моделирование»,
3) ключевые слова, с точностью до порядка слов повторяющие название рубрики («движение вращательное»),
4) словосочетания, почти дословно повторяющие название её родительской рубрики (термин «движение возмущенное»).

Для упрощения модели мы предполагаем, что ключевые слова являются независимыми, т.е. появление одного из них не влечет за собой (с большой вероятностью) появление другого из некоторого списка.

Нормализация рубрик верхнего уровня по числу документов

Неравномерность распределения потока документов по выпускам РЖ вызывает неравномерное распределение их по рубрикам 1-го уровня. Разрыв может достигать 100 и более раз. Поэтому для определения принадлежности ключевого слова области знаний на основе статистических данных необходимо ввести весовые

коэффициенты для связей «документ – ключевое слово» так, чтобы количество «взвешенных» рефератов во всех областях было примерно одинаковым.

Для каждой ветви Рубрикатора введем *вес*, обратно пропорциональный количеству документов в ней. В зависимости от выбора коэффициента можно сделать либо максимальный, либо минимальный вес равным 1; в последнем случае в качестве веса могут быть взяты целые числа.

Для того чтобы снизить влияние часто употребляемых ключевых слов добавим признак, характеризующий обратную встречаемость термина в документах, который также может быть нормализован.

Задел для решения проблем построения тезауруса

Разбиение массива ключевых слов на кластеры с выделением «канонического» представления и разнесение их по рубрикам может быть использовано при решении проблемы построения поискового тезауруса по тематике ВИНИТИ.

Построенные в настоящее время кластеры взаимно заменяемых словоформ могут быть дополнены синонимами и переводами термина на другие языки. Они вносятся в кластер только в случае полного соответствия их исходному понятию. Внесение в кластер синонимов и перевода – ручная работа специалистов.

При определении связей между кластерами сложность заключается в установлении их автоматизировано, с минимальным участием экспертов. Возможно, это удастся сделать по наличию уточняющих слов. С определенной долей вероятности можно считать, что добавление в словосочетание ещё одного слова образует либо более узкий термин («граф» – «*полный* граф»), либо – другое понятие («полный граф» – «полный *двудольный* граф»). Для установления иерархии между терминами можно выдвинуть следующую гипотезу: для двух терминов, приписанных к иерархически связанным рубрикам и отличающихся только наличием уточняющих слов, термин, отнесенный к рубрике более высокого уровня, охватывает другой, т.е. является более широким понятием. Для этой операции могут понадобиться все словосочетания, входящие в кластер.

Расширение синонимии за счет допущения, что два синонима одного и того же термина могут не быть синонимами между собой, вызовет замену имеющегося представления кластеров на аппарат ссылок между ключевыми словами – синонимами. В результате получим граф отношений синонимии. Для каждого термина список его синонимов – «звезда», имеющая центром данный термин. Группы синонимов могут пересекаться. В развитие этого подхода можно предложить «раскрасить» ребра графа шифрами тематик или областей наук, а также «взвесить» связи по степени близости понятий. Это также позволит отчасти решить проблему омонимов. Кроме того, признаком терминов–омонимов является использование их в предметном описании рубрик разных ветвей Рубрикатора в качестве основных терминов.

Список литературы

1. *Андронова М.Б., Ефременкова В.М.* Определение тематической направленности журналов на основе Классификационных систем отраслей знания. // НТИ. Сер. 1. М.: ВИНИТИ, 2001. № 7. С. 15-23.
2. *Андронова М.Б., Ефременкова В.М., Хитрово Н.Г.* Состояние и проблемы формирования информационного массива по тематике "Техника ориентации в трехмерном пространстве" // НТИ. Сер. 1. М.: ВИНИТИ, 1999. № 7. С. 27-34.
3. Рубрикатор информационных изданий ВИНИТИ. – Т.1, Т. 2. // М.: ВИНИТИ, 1999. 448 С.
4. *Федорец О.В.* Поиск по сходству в реляционной базе данных: статистический подход к хешированию библиографических записей // НТИ. Сер. 2. М.: ВИНИТИ, 2005. № 1. С. 9-21.