Формирование и ведение тезауруса в составе посредника между пользователями и сетью электронных библиотек

Казаков Е.Н.

Всероссийский научно-исследовательский технический информационный центр
Москва, Россия
E-mail:postmaster@vntic.org.ru

Электронные библиотеки (ЭБ) представляют собой электронную информационную среду, предназначенную для взаимодействия с самыми широкими слоями населения. Нереально требовать от пользователя изучения особенностей доступа к каждой отдельной ЭБ при том их количестве и разнообразии, которое предоставляет Интернет. Наиболее перспективный, но труднореализуемый путь – создать посредник, который фиксирует особенности каждой ЭБ или отдельной коллекции и преобразует конкретный запрос пользователя в серию запросов, учитывающих эти особенности.

Принципы функционирования и структура посредника разрабатываются под руководством Л.А. Калиниченко в Институте проблем информатики Российской академии наук (ИПИ РАН).

Для отображения лексики и семантики различных коллекций и для интеллектуальной помощи пользователям при взаимодействии с ЭБ в состав посредника целесообразно ввести тезаурус. Проблема использования тезауруса в составе посредника при подключении коллекций и при взаимодействии с пользователями рассмотрены в [1]. Лексика и семантика тезауруса посредника должна соответствовать лексике и семантике тех коллекций, которые зарегистрированы в посреднике. Учитывая разнообразие тематики коллекций, необходимо использовать политематический тезаурус.

Поскольку многие коллекции, зарегистрированные в посреднике, могут иметь собственные тезаурусы, то возникает задача интеграции тезаурусов. Следует отметить, что задача интеграции тезаурусов возникает именно в связи с формированием единого тезауруса посредника. Ранее эта задача не возникала, поскольку каждая автономная информационная система имела собственный тезаурус.

Возможны разные стратегии интеграции тезаурусов. Если в распоряжении создателей посредника имеется готовый политематический тезаурус, то его можно на начальном этапе сделать нормативным и сравнивать с ним тезаурусы или представительные тексты всех подключаемых к посреднику коллекций.

На основе этих сравнений определяются: лексика, совпадающая с нормативным тезаурусом (НТ) и дополнительная лексика, специфичная для каждой коллекции, Часть этой дополнительной лексики используется для пополнения НТ и формирования в итоге интегрированного тезауруса (ИТ) посредника.

Если готовый политематический тезаурус отсутствует, то интегрированный тезаурус посредника формируется как объединение тезаурусов всех подключенных к посреднику коллекций.

В докладе более подробно рассматривается проблема интеграции на базе нормативного политематического тезауруса. Интеграция тезаурусов включает интеграцию лексики и интеграцию семантических отношений. Интеграция лексики рассматривается как на уровне словоформ, так и на уровне словосочетаний.

1. Интеграция лексики на уровне словоформ

С целью интеграции лексики на уровне словоформ для НТ и тезаурусов коллекций составляются алфавитные словари грамматически нормализованных словоформ, входящих в лексические единицы (ЛЕ). Существительные представляются в именительном падеже единственного числа (кроме существительных, не имеющих единственного числа), прилагательные – в именительном падеже единственного числа мужского рода.

Для каждого алфавитного словаря формируются списки словоформ, совпадающих и не совпадающих со словоформами НТ.

Для количественной характеристики соотношения лексики НТ и лексики коллекций предлагаются:

- коэффициент совпадения словоформ ($К_с$);
- коэффициент покрытия нормативного тезауруса словоформами i-ой коллекции ($К_п^{нт,i}$);
- коэффициент покрытия словоформ i-ой коллекции словоформами нормативного тезауруса ($К_п^{i}$);

Пусть

$S_{нт}$ – множество словоформ НТ,

S_i – множество словоформ в алфавитном словаре i-ой коллекции,

$S_{i,нт}$ – множество словоформ i-ой коллекции, совпавших со словоформами НТ.

Обозначим $|S|$ - мощность множества S.

Введем $К_с = \dfrac{2|S_i, _{нт}|}{|S_{нт}| + |S_i|}$

Очевидно, что $К_с = 1$, если $S_{нт} = S_i$,

$К_с = 0$, если $|S_{i,нт}| = 0$.

Первая Всероссийская научная конференция
ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,
ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ
19 - 21 октября 1999 г., Санкт-Петербург

$K_п^{нт,i} = \frac{|S_{i,\,нт}|}{|S_{нт}|}$, причем $K_п^{нт,i} = 1$, если $S_{нт} = S_i$,

$K_п^{нт,i} = 0$, если $|S_{i,нт}| = 0$.

$K_п^{i} = \frac{|S_{i,\,нт}|}{|S_i|}$, причем $K_п^{i} = 1$, если $S_{i,нт} = S_i$

$K_п^{i} = 0$, если $|S_{i,нт}| = 0$.

В случае достаточно полного политематического тезауруса и узкотематической i-ой коллекции

$K_п^{нт} \to \frac{|S_i|}{|S_{нт}|} << 1$, а $K_п^{i} \to \frac{|Si|}{|Si|} = 1$.

Если $K_п^{i}$ существенно меньше 1, это означает, что тематика коллекции не совпадает с тематикой, отраженной в НТ и, следовательно, лексика этой коллекции должна использоваться для пополнения НТ.

Дополнительная лексика i-ой коллекции относительно НТ включает словоформы i-ой коллекции, не совпавшие со словоформами НТ, и определяется выражением $S_i \backslash S_{i,нт.}$

Дополнительная лексика посредника формируется на основе объединения дополнительной лексики всех зарегистрированных коллекций, т.е. образует множество

$$\bigcup_{i=1}^{n} S_i \setminus S_{i,\,нт},$$

причем каждой словоформе приписывается суммарная частота употребления и количество коллекций, в которых она встречается.

Для дополнительной лексики посредника целесообразно построить ранговые распределения первого и второго рода.

Чтобы построить ранговое распределение первого рода, необходимо все словоформы упорядочить по убыванию суммарной частоты употребления во всех зарегистрированных коллекциях.

Словоформе с максимальной частотой, приписывается ранг первого рода 1, следующей – 2 и т.д.

Чтобы построить ранговое распределение второго рода, необходимо все словоформы упорядочить по убыванию количества вхождений во множества дополнительной лексики всех зарегистрированных коллекций. Словоформе с максимальным количеством вхождений, приписывается ранг второго рода 1, следующей по порядку – 2 и т.д.

Ранговое распределение первого рода отдает предпочтение наиболее употребительным словоформам, а ранговое распределение второго рода выделяет те словоформы, которые входят в различные коллекции. Ранговое распределение второго рода имеет смысл строить при достаточно большом количестве зарегистрированных коллекций.

Анализируя ранговые распределения, можно для каждого назначить некоторый пороговый ранг. В этом случае лексику нормативного тезауруса пополнят все словоформы, ранг которых ниже порогового значения.

2. Интеграция лексики на уровне словосочетаний

Сопоставляя различные тезаурусы на уровне словосочетаний, можно заметить, что в некоторых тезаурусах сходные понятия выражаются более длинными словосочетаниями, тогда как в других передаются комбинацией более коротких словосочетаний. Например: понятие одного тезауруса “автоматизированные системы управления” может быть в другом тезаурусе представлено комбинацией лексических единиц (ЛЕ) “автоматизированные системы” и “системы управления”.

Для количественной характеристики соотношений лексики тезаурусов на уровне словосочетаний необходимо поставить все тезаурусы в равные условия, не зависящие от длины словосочетаний. С этой целью для каждого тезауруса формируется алфавитный словарь минимальных словосочетаний, в который включаются двухсловные словосочетания, входящие в ЛЕ из двух и более слов.

Количественные характеристики соотношений лексики на уровне словосочетаний определяются по аналогии со словоформами.

Пусть

$C_{нт}$ – множество минимальных словосочетаний в нормативном тезаурусе,

C_i – аналогичное множество в тезаурусе i-ой коллекции,

$C_{i,нт}$ – множество минимальных словосочетаний i-ой коллекции, совпавших с аналогичным множеством нормативного тезауруса.

Тогда коэффициент совпадения лексики на уровне словосочетаний ($K_{сс}$) определится выражением

$$K_{сс} = \frac{2|C_{i,\,нт}|}{|C_{нт}| + |C_i|},$$

причем $K_{сс}=1$, если $C_{нт}=C_i$,

$K_{сс} = 0$, если $|C_{i,нт}| = 0$.

Коэффициент покрытия лексики нормативного тезауруса словосочетаниями i-ой коллекции ($K_п^{нт,i}$)

$$K_{пс}^{нт,i} = \frac{|C_{i,\,нт}|}{|C_{нт}|}$$

причем $K_{пс}^{нт,i} = 1$, если $C_{нт} = C_i$

$K_{пс}^{нт,i} = 0$, если $|C_{i,нт}| = 0$.

Коэффициент покрытия лексики i-ой коллекции минимальными словосочетаниями нормативного тезауруса ($K_{пс}^{i}$):

$$K_{пс}^{i} = \frac{|C_{i,\,нт}|}{|C_i|}$$

причем $K_{пс}^{i} = 1$, если $C_i = C_{нт}$

$K_{пс}^{i} = 0$, если $|C_{i,нт}| = 0$.

Значения коэффициентов позволяют оценивать необходимость интеграции. Так, если $K_{пс}^{i}$ существенно меньше единицы, то необходимо пополнение НТ, поскольку значительная часть лексики i-ой коллекции в нем отсутствует.

Технология интеграции на уровне словосочетаний аналогична технологии на уровне словоформ.

Для объединенного словаря двухсловных словосочетаний, не входящих в нормативный тезаурус,

$$\bigcup_{i=1}^{n} C_i \setminus C_{i,\,\text{НТ}}$$

строятся ранговые распределения первого и второго рода, назначаются пороговые ранги и нормативный тезаурус пополняют теми словосочетаниями, ранг которых не превышает пороговое значение. Словосочетания включаются в нормативный тезаурус в той форме, в которой оно встречается в тезаурусах большей части коллекций.

Следует отметить, что перенесение в посредник из конкретных коллекций лексики сначала НТ, а впоследствии интегрированного тезауруса (ИТ), а также фиксация дополнительной лексики позволяют упростить представление лексики, поскольку лексика интегрированного тезауруса не повторяется в многочисленных коллекциях и играет роль общего знаменателя, вынесенного за скобки коллекций.

Целесообразно подчеркнуть, что дополнительная лексика может использоваться для поиска и поддержки пользователей сразу после подключения коллекций к посреднику, не дожидаясь ее ввода в ИТ.

В связи с использованием ранговых распределений можно заметить, что ранговые частотные распределения ЛЕ (словоформ и словосочетаний) подчиняются закону Ципфа:

$$p_r = \frac{k}{r}$$

p_r - вероятность появления в тексте ЛЕ с рангом r,

r - ранг ЛЕ,

k - некоторая константа.

Закон Ципфа выражает универсальное свойство всех естественных языков: основную часть текста образует сравнительно небольшое число употребляемых слов и словосочетаний.

Опыт формирования политематического словаря ВНТИЦ показывает, что примерно половина наиболее частотных ЛЕ образуют примерно 90 % текста. Поэтому при формировании и пополнении тезауруса на основе анализа текста множество низкочастотных ЛЕ (с частотой 1 ÷ 3), так называемый “отстойник”, можно не включать в тезаурус, сохраняя эти ЛЕ в массиве дополнительной лексики с целью использования при поиске и взаимодействии с пользователем.

Опыт показывает, что в рамках одного естественного языка в сфере научно-технической информации в каждой тематической области существует ограниченное число терминов и понятий, которыми оперирует научно-техническое сообщество. Это количество понятий практически не превышает 30 - 40 тысяч, за исключением химии и биологии.

Отсюда следует возможность эффективного использования в научно-технической сфере политематического тезауруса, полученного на основе обработки лексики одного вида документов для всех видов документов данной сферы.

3. Интеграция семантических отношений

Для интеграции семантических отношений необходимо в тезаурусах коллекций выявить все лексические единицы (ЛЕ), присутствующие в нормативном тезаурусе и имеющие семантические отношения, отсутствующие в НТ. С этой целью для каждого из тезаурусов формируется алфавитный список ЛЕ, каждая из которых имеет хотя бы одно семантическое отношение (связь) с другой ЛЕ. В этом списке фиксируется частота употребления ЛЕ, виды связей и перечень ЛЕ, с которыми есть связь каждого вида.

Алфавитный список каждой коллекции сравнивается с алфавитным списком НТ.

В результате сравнения все ЛЕ алфавитного списка каждой коллекции разделяются на три множества

$\left(L_i^{=}\right)$ - ЛЕ i-ой коллекции, точно совпадающая с ЛЕ нормативного тезауруса

$\left(L_i^{\sim}\right)$ - ЛЕ i-ой коллекции, частично совпадающие $L_i^{=}, L_i^{\sim}, L_i^{\neq}$

$\left(L_i^{\neq}\right)$ - ЛЕ i-ой коллекции, не совпадающие с ЛЕ нормативного тезауруса

Для каждой ЛЕ из $L_i^{=}$ определяются ЛЕ, имеющие с ней связи и виды связей. Среди этих ЛЕ и связей выделяются ЛЕ, не принадлежащие $L_i^{=}$, или вид связи, отличающийся от связей нормативного тезауруса. Именно эти ЛЕ и виды связей используются для пополнения нормативного тезауруса.

В целом для посредника множества $L_i^{\sim}$ и $L_i^{\neq}$ объединяются и образуют множества $L^{\sim}$ и $L^{\neq}$. ЛЕ и соответствующие им виды связей из этих множеств упорядочиваются либо по убыванию суммарной частоты употребления, либо по убыванию числа коллекций, в которые они входят с соответствующим присвоением ранга.

Затем выбирается пороговое значение ранга и те ЛЕ и виды связей, которые не превышают порога включаются в нормативный тезаурус. Специальный анализ требуется для ЛЕ из множества $L^{\sim}$. Поскольку ЛЕ из этого множества частично совпадают с ЛЕ нормативного тезауруса, то необходимо принять решение: либо сохранять ЛЕ нормативного тезауруса, дополнив связи, либо ввести ЛЕ из $L^{\sim}$, либо ввести ЛЕ, отличающуюся и от ЛЕ нормативного тезауруса и от ЛЕ из $L^{\sim}$.

При интеграции связей необходимо учитывать, что синонимические связи удовлетворяют условиям эквивалентности, а иерархические связи условиям транзитивности. Если в нормативном тезаурусе лексические единицы b_1 и b_2, а в множествах являются синонимами, т.е. $b_1 = b_2$, а в множествах $L^{=}$, $L^{\sim}$, $L^{\neq}$ $b_2 = b_3$, то происходит формирование цепочки синонимов $b_1 = b_2 = b_3 = b_1$.

Если в нормативном тезаурусе лексические единицы c_1 и c_2 имеют отношение c_1 выше c_2, а в множествах $L^{=}$, $L^{\sim}$, $L^{\neq}$ c_2 выше c_3, то фиксируются связи c_1 выше c_2 выше c_3. Если же в НТ встречается иерархия, на нижнем уровне которой есть лексическая единица c , а в множествах $L^{=}$, $L^{\sim}$, $L^{\neq}$ есть иерархия, вершиной которой является c , то вся иерархия переносится в НТ.

Необходимо подчеркнуть, что все описанные процедуры могут использоваться не только при формировании интегрированного тезауруса, но и в процессе ведения ИТ, когда его нужно модифицировать либо в случае изменения тезаурусов зарегистрированных коллекций, либо в случае подключения новых коллекций к посреднику.

В случае отсутствия политематического тезауруса формирование интегрированного тезауруса посредника можно осуществлять с помощью описанных процедур с нуля.

В этом случае сначала по описанной методике сравниваются тезаурусы, относящиеся к одной тематике, и среди них в качестве представительного тезауруса выбирается наиболее полный, т.е. максимально покрывающий лексику всех сравниваемых тезаурусов. Затем в качестве ИТ можно при-

нять тезаурус, являющийся объединением представительных тезаурусов по различным тематическим областям и впоследствии по описанным процедурам осуществлять интерактивное ведение и пополнение ИТ.

Литература

1. КАЗАКОВ Е.Н. Использование политематического тезауруса для поддержки пользователей при взаимодействии с сетью электронных библиотек // Аналитический вестник. - М. : ВНТИЦ, 1999, № 1, с. 17 - 21.