О возможности содержательного индексирования документов ключевыми словами

(по материалам эксперимента, проведенного в РНБ на базе комплекса "Охрана окружающей среды")

Н.П. Никольцева,
О.А. Седышева

Разработка лингвистического обеспечения ЭК РНБ потребовала определения оптимального состава ИПЯ и принципов их использования. Частью комплекса работ в этом направлении явился эксперимент, одной из задач которого было рассмотрение возможности использования помимо традиционного для РНБ аппарата индексирования (языки ПР и ББК) дополнительного языка, например, КС. Эксперимент проводился в 1995-96 гг. на основе БД комплекса литературы "Охрана окружающей среды".

Развитие иформационных систем on-line и переход от традиционных систем к машиночитаемым дали возможность работать с ЕЯ. Появились системы, в которых сочетаются возможности контролируемого и неконтролируемого языка /10,11,12/. Значительное внимание уделяется сравнению этих языков, исследованию вопросов контроля словаря- нужен ли он вообще и в какой степени, какой баланс должен быть между контролируемым словарем и использованием ЕЯ.

Достаточно полное сравнение контролируемого и неконтролируемого языка проведено в работах /6,13/. Остановимся только на самых важных моментах.

В качестве основных преимуществ ЕЯ в литературе приводятся следующие:

1. подробность, т.е. детальность индексирования
2. исчерпывающий характер отражения содержания
3. оперативное обновление лексики
4. использование авторской терминологии

К преимуществам контролируемого языка относятся возможности:

1. контроля синонимии
2. определения омографов
3. представления термина со всевозможными связями
4. предоставления примечания о применении
5. применения грамматического и логического синтаксиса, исключающего ошибки поиска из-за неправильных комбинаций терминов, как это присутствует при неконтролируемом языке.

ЕЯ имеет ряд преимуществ по сравнению с контролируемым языком, и наоборот. В "смешанных" системах, использующих оба языка, возможности, присущие каждому из языков, дополняют друг друга.

Существование того или иного типа систем в различных странах определяется:

- исторически сложившейся библиотечной практикой и накопленными БД /7, 11, 12/
- развитием компьютерных технологий.

Появились системы с автоматическим индексированием и поиском/9, 15/. Автоматическая обработка текстов ЕЯ с целью выявления их содержания была и остается одним из самых важных направлений исследований в крупнейших информационных центрах мира.

Например, институт научной и технической информации национального центра научных исследований Франции разрабатывает лингвинистический подход, включающий модель структурированного индексирования/14/, Японская фирма NTT разработала систему автоматического выбора ключевых терминов из научных статей на японском языке.

Широкое развитие получили полнотекстовые БД, где в ПОД включаются все значимые слова текста. Но при таком подходе, когда для индексирования используется только автоматическая обработка документа КС, при поиске (учитывая "причуды ЕЯ") выдача информации часто оказывается неадекватной.

Подобные методы обработки документов ЕЯ пока еще не обеспечивают эффективного, независимого от тематической области, строгого и точного анализа текста документа /8,15/.

В настоящее время чаще встречаются системы с комбинированным использованием контролируемого и неконтролируемого языков, причем последний состоит из КС, автоматически выбираемых из БЗ (т.е. заглавий, рефератов, аннотаций, серий и т.д.).

Возможность содержательного индексирования документов КС в дополнение к автоматической выборке КС из БЗ и предметному индексированию рассматривалась в ходе проведенного в РНБ эксперимента. Такой подход предоставляет дополнительное средство для описания документа на ЕЯ при помощи метода свободного координатного индексирования.

Метод координатного индексирования базируется на представлении о том, что основное смысловое содержание документа может быть с достаточной степенью точности и полноты выражено набором КС, содержащихся в индексируемом тексте /4/. Свободное координатное индексирование означает индексирование КС, выбранными непосредственно из полного текста документа и представленными в ПОДе в терминологии автора без нормализации, с минимальным контролем над лексикой и без учета того, какие КС уже использовались ранее для индексирования таких же или близких по смыслу документов. (В отличие от методики индексирования с применением тезауруса, где ПОД составляется на основе КС документа, но представленных в виде устоявшихся понятий, сверенных с тезаурусом). При свободном индексировании словарь КС создается вместе с формированием базы.

Остановимся более подробно на понятии "КС" /1, 3, 4, 5/.

Ключевыми словами в традиционном понимании называются полнозначные слова, устойчивые сложные и сложносокращенные слова и терминологические словосочетания, несущие смысловую нагрузку в текстах документов.

В качестве КС могут выступать слова (унитермы), терминологические словосочетания, фразы (в исключительных случаях), аббревиатуры, численные характеристики, хронологические данные, имена собственные, символические обозначения.

Основное требование, которое предъявляется к выбору КС - строгая определенность значения.

Для каждой предметной области в первую очередь следует выделять специальные (специфические) термины, употребляемые только в данной отрасли знания, являющиеся как бы ее опознавательными знаками.

В качестве КС не рекомендуется использовать:

1. Общеупотребительные термины типа - проблема, значение, метод, принцип, свойства, уравнение.
   Включение таких терминов в ПОД возможно только в сочетании с другими словами, сужающими их значение.

   Например: Принцип свободного оседания

   Органолептические свойства
   Уравнение турбулентной диффузии

2. Служебные слова (частицы, предлоги, союзы, междометия).
3. Термины, частота встречаемости которых в данном документе мала, если они не являются узкоспецифическими.

В качестве КС могут выступать:

1. *Термины - слова (унитермы)*
   Могут быть существительными (нарицательными и собственными).

   Например: кальций

   охрана
   Казань

   Прилагательные, наречия, числительные, местоимения, которые сами по себе могут выступать в роли КС, предлагаем употреблять только в словосочетаниях.

2. *Термины - словосочетания*

   Например: гигиеническое нормирование

   сточные воды
   рыбные запасы
   Арктический бассейн

3. *Термины-предложения*
   Термины-предложения следует оставлять, не деля их на словосочетания, лишь в том случае, когда такое разделение может привести к потере смысла исходного предложения. Например такими КС могут быть различные команды.
4. *Численные характеристики*
   Численные характеристики (диапазоны температур, давлений и т.д.) следует давать со словесной формулировкой.

   Например: диапазон температур 373- 453 К

5. *Хронологические данные*

   Например: 1937 -1987 гг., 20в и т.д.
   Возможно включение в ПОД словесного обозначения периода.

   Например: средневековье.

6. *Символические обозначения*

   Например: Sr-90

7. *Аббревиатуры*
   Аббревиатуры используются в том случае, когда краткая форма вытеснила полное наименование предмета и сокращение чаще используются в терминосистемах.

   Например: ПДК, ЭВМ.

Методика выборки и форма представления КС в традиционных системах были ориентированы на дальнейшую нормализацию КС и создание карточных тезаурусов. В случае ЭК и с учетом того, что КС будут представлены в ПОДах в ненормализованном виде, некоторые требования, предъявляемые к форме представления КС, как нам кажется, становятся необязательными. К таким требованиям относятся:

- употребление устойчивых словосочетаний (критерии устойчивости подробно рассмотрены в работе/5/)
- устранение синонимии, омонимии и полисемии /1/
- представление КС в единообразной грамматической форме /1/.

В традиционных системах при формулировке КС в ПОДах в основном используют унитермы, сохраняя в ряде случаев устойчивые словосочетания. При унитермном подходе, путем свободного манипулирования элементами поисковых образов, можно обеспечить любую глубину и детальность индексирования, увеличить число точек доступа к искомым документам, но при этом может уменьшиться точность поиска за счет появления паразитных и неадекватных лексических конфигураций.

Например: моделирование распространения - распространение моделирования

программа расчета - расчет программ

С нашей точки зрения, для целей ЭК целесообразнее использовать словосочетания, причем не только устойчивые, но и неустойчивые, так как это является способом:
а) устранения информационного шума
б) устранения полисемии и омонимии

Например, если разделить неустойчивое словосочетание:

уравнение турбулентной диффузии
на унитермы, то получаются следующие термины
турбулентный
диффузия
уравнение - вообще в список КС не попадает, так как это общенаучный термин. При таком списке КС индивидуальность документов пропадает, так как турбулентной может быть не

только диффузия, но и поток, процесс и т.д.; диффузия - ионно-обменной, ламинарной и т.д., без "уравнения" даже словосочетание "турбулентная диффузия" может оказаться описанием физического явления, тогда как в данном случае речь идет исключительно о математических выкладках.

Проблема синонимии. В случае свободного индексирования, где в ПОД включаются КС в терминологии автора, избежать синонимии практически невозможно. Но и при индексировании одного документа (особенно это касается сборников, посвященных одной проблеме) в ПОДе могут оказаться термины-синонимы, устранить которые индексатор обязан.

Например: ситаллы

пирокерамы
Из этих терминов может существовать в ПОДе только один.

Формулировка КС в ЭК во многом зависит от поисковых возможностей ИПС. Например, для ведения БД комплекса литературы "Охрана окружающей среды" был использован пакет прикладных программ CDS/ISIS/M, где с помощью инвертированного файла можно создать для каждой записи фактически неограниченное число точек доступа, причем в качестве поисковых можно выбирать любые поля, подполя БО документа, извлекать отдельные слова или целые фразы. С помощью Булевых операторов "и", "или", "не" возможно:

1. выделение тех записей, в которых встречаются несколько заданных терминов одновременно
2. проводить поиск по словосочетаниям, скомбинированным из унитерм, причем можно указать позицию слов относительно друг друга
3. проводить поиск по усеченным или точно заданным терминам, т.е. можно не учитывать число и падеж существительных. Поэтому требование единообразного грамматического представления КС в случае ЭК также становится необязательным.

Таким образом, мы считаем возможным использовать в качестве КС унитермы и любые словосочетания (устойчивые и неустойчивые), т.е. то, что будет отражать индивидуальность документа, действительно являться его КС, а также фразы, если разделение их на словосочетания приведет к потере адекватности отражения содержания документа. КС выбираются из документа в терминологии автора и в дальнейшей обработке не нуждаются.

Рассмотрим, из каких фрагментов текста следует выбирать КС. Полноценное индексирование обычно проводится не по полному тексту документа, а по его значимым фрагментам (титульный лист, аннотация, реферат, предисловие, оглавление, библиография и т.д.). Иногда необходим просмотр всего текста документа. Подробно на этих вопросах в рамках данной статьи останавливаться не будем. Отметим лишь, что фрагментарный анализ документа с целью выборки КС производится также как и анализ для присваивания ПР.

Вернемся к вопросу о целесообразности использования КС, если имеется развитый аппарат для индексирования документа - язык ПР (язык ББК в данной статье не рассматривается). Нужно ли дополнять ПОД КС?

Результаты эксперимента показали , что целесообразно использовать КС для дополнительного раскрытия содержания документа на более глубоком уровне, т.е. использовать ПР для описания основных предметов документа и их аспектов, КС- для их дальнейшей детализации а также описания побочных тем документа). Напомним, что глубина координатного индексирования определяется степенью полноты отражения в ПОДе всех наиболее важных специфических понятий, рассматриваемых в этом документе и выражаемых в его ПОДе при помощи наиболее специфических КС/4/, и не измеряется их числом.

В ходе эксперимента было заиндексировано 745 документов. Каждому документу присваивалось от 5 до 35 КС и от 1 до 11 ПР. Время, затраченное на индексирование одного документа составляло от 5 до 30 мин.

Для определения соотношения ПР и КС список КС в приведенных ниже примерах не отредактирован, т.е. не исключены термины, совпадающие лексически с КС из ПР и БО (КС - ненормализованная лексика, ПР - взяты произвольно, без привязки к ПК РНБ).

Сравнение индексирования документов ПР и КС проводилось по полноте отражения предметов и их аспектов в документе. Лексическое сравнение проводить не корректно в связи с тем, что язык ПР- контролируемый, а язык КС - неконтролируемый.

Полученые результаты можно разделить на три группы:

1. КС и ПР идентичны по полноте отражения содержания документа. Таких случаев - 9%, причем число ПР, присвоеных на один документ, было от 2 до 11, а КС - от 3 до 35. Следует отметить, что наблюдалось совпадение и лексики ПР и КС.

   Пример 1. Временные методические указания по химическому анализу атмосферного воздуха с отбором проб на твердые пленочные сорбенты/Гос.ком.СССР по гидрометеорологии и контролю природной Среды; Под ред. к.т.н. Н.Ш.Вольберга.-Л.:Гидрометеоздат, 1982.- 34с.

   КС

   химический анализ
   атмосферный воздух
   отбор проб
   твердые пленочные сорбенты
   сорбционные трубки
   стеклянные гранулы
   двуокись серы
   двуокись азота
   сероводород
   фтористый водород
   хлористый водород
   метилмеркаптан
   фенол
   диметиламин
   сероуглерод

   ПР

1. Атмосферной воздух - Химический анализ
2. Твердые пленочные сорбенты
3. Двуокись серы - Определение
4. Двуокись азота - Определение
5. Сероводород - Определение
6. Хлористый водород - Определение
7. Фтористый водород - Определение
8. Метилмеркаптан - Определение
9. Диметиланилин - Определение
10. Фенол - Определение
11. Сорбционные трубки

Пример 2. Ковшарь А.Ф. Заповедник Аксу-Джабаглы.- Алма-Ата: Кайнар, 1982.- 160 с.

КС

1. заповедники
2. Аксу-Джабаглы
3. Казахская ССР
4. туристические маршруты

ПР

1. Аксу-Джабаглы-Заповедник
2. Туристские маршруты- Казахская ССР

2. В 11% проиндексированных документов КС лишь частично увеличивают полноту индексирования. Документам присваивалось большое число ПР (более 3).

Пример 3. Ильницкий А.П. и др. Канцерогенные вещества в водной среде/Ильницкий А.П., Королев А.А., Худолей В.В.; Ин-т геохимии и аналит. химии им. В.И.Вернадского.-М.:Наука, 1993.-219 с.

КС

канцерогенные вещества
водная среда
онкологические заболевания
водоемы
сточные воды
ливневые воды
талые воды
промышленные отходы
опухоли рыб
опухоли моллюсков
загрязнение
биологический мониторинг
диагностический мониторинг
прогностический мониторинг
питьевая вода

очистка
полициклические ароматические углеводороды
нитрозосоединения
хлорорганические соединения
нитраты
нитриты
бензопирен
нитроздиэтиламин

ПР

1. Канцерогенные вещества - Содержание в водоемах
2. Питьевая вода - Очистка
3. Сточные воды - Очистка
4. Водоемы - Загрязение канцерогенными веществами - Мониторинг
5. Рыбы - Опухоли
6. Моллюски - Опухоли

Пример 4. Справочник общественного инспектора по охране природы Эстонской ССР/М-во лесного хоз-ва и охрана природы ЭССР.-Таллин.: Валгус, 1972.- 290с.

КС

охрана природы
Эстонская ССР
природные ресурсы
полезные ископаемые
вода
почвы
растительность
фауна
ландшафты
мелиорация
атмосфера
водоемы
загрязнение
очистные сооружения
сточные воды
очистка
биологическая очистка
обеззараживание
леса
охотничьи хозяйства
рыболовство
плавучие средства
загрязнение
атмосферный воздух

ПР

1. Охрана природы - Эстонская ССР - Справочники
2. Полезные ископаемые - Охрана
3. Почвы - Охрана
4. Растения - Охрана
5.Ландшафты - Охрана
6. Атмосфера - Охрана
7. Водоемы - Охрана
8. Леса - Пожарная безопасность
9. Животные - Охрана
10. Мелиорация
11. Сточные воды - Очистка

3. В 80% обработанных документов полнота индексирования ПР и КС была различной. Документу давались ПР, охватывающие предметы произведения в целом, а КС более детально отражали эти предметы и их аспекты с учетом специфических терминов. ПОДы в среднем содержали 10- 15 КС и 1- 3 ПР.

Пример 5. Стоберски Д. Влияние промышленных предприятий на загрязненность воздушной среды города. Автореф. ... канд.техн.наук.-М., 1989.-16 с.

КС

промышленные предприятия
загрязненность
города
воздушная среда
атмосферный воздух
вредные вещества
охрана
математическое моделирование
внутризаводской автотранспорт
законодательные акты
промышленные зоны
санитарно-защитные зоны

ПР

Города- Воздушная среда - Загрязнение

Пример 6. Родные просторы: Сб. статей и очерков о природе родного края/Всерос. о-во содействия охране природы и озеленению населенных пунктов.- Воронеж: Кн. изд., 1961. -55 с.

КС

природа
охрана
Дон
Хоперский заповедник
Хреновский бор дикорастущие бобовые травы
бобры

жабы
земноводные
ценные животные
лесные насекомые
фенологические наблюдения
зеленый патруль
Воронежская область

ПР

Природа - Охрана - Воронежская область

Пример 7. Чернобережский Ю.М. и др. Основы микробиологи и химии воды: Уч. пособие/Чернобережский Ю.М., Николаев А.Н., Вольф И.В.; М-во высш. и сред. спец. образования РСФСР, Ленингр. лесотехн. акад. -Л.: ЛТА, 1988. - 84с.

КС

природные воды
сточные воды
реки
озера
физико-химический состав
примеси
загрязнение
очистка
коагуляция
гетерокоагуляция
флокуляция
флотация
молекулярная адсорбция
ионный обмен
электроанализ
гиперфильтрация
микроорганизмы
бактерии
грибы
водоросли
простейшие
рост

ПР

Вода - Микробиология
Вода - Химия

Такие результаты показали, что если полнота индексирования КС и ПР одинакова, то нет необходимости в КС. Но, как правило, ПР по сравнению с КС являются "обобщающим" описанием содержания документа. В таких случаях дополнение ПР КС дает значительный эффект с точки зрения полноты индексирования.

Существуют и другие доводы "за" использование КС.

Возможны случаи, когда некоторые побочные темы документа не описываются ПР. Например, это часто бывает при индексировании материалов конференций, симпозиумов, где иногда можно встретить тезисы докладов, отдаленно напоминающих основную тематику. С большой долей вероятности такие побочные темы не будут отражены ПР, но могут быть описаны КС.

Например: VI Всесоюзный симпозиум по химии неорганических фторидов. Тезисы докладов. 21-23 июля 1981 г. Новосибирск, 1981.-299 с. Основную тему данного документа можно описать ПР: Неорганические фториды - Химия. Однако в тезисах есть доклад, посвященный изучению тепломассообмена в барабанных печах. Следовательно, данная тема (оборудование химических производств) не будет отражена в ПОДе. Для того, чтобы подобная информация не была потеряна для пользователя, можно дополнить ПОД описанием этой темы КС.

Из анализа результатов эксперимента следует, что целесообразно проводить дополнительное индексирование КС при условии, что КС не будут повторять ПР, для увеличения полноты и детальности индексирования, для отражения побочных тем документа. В ПОДе будут находится ПР (нормализованная лексика) и КС в редакции автора. Такой подход удобен для всех категорий читателей: для тех, кто примерно знает, что хочет найти (для этого удобно использование ПР) и для тех, кто может спросить что-то очень специфическое (по КС). Особую ценность список ненормализованных КС может приобрести на последнем этапе поиска, когда на запрос найден какой-то массив документов (поиск по ПР). Просмотрев списки КС найденных документов, можно сразу определить, нужна ли книга или нет, не получая ее на руки.

Использование языка КС может быть особенно интересно для областных, профильных, узкоспециализированных библиотек, библиотек с неразвитым аппаратом ПР или специальных БД, например по краеведению, где часто источниками информации могут служить газетные публикации, сборники или монографии, которые посвящены другим предметам, но содержат данные, интересные для краеведов. В этих случаях индексирование КС будет являться очень ценным дополнением к другим средствам индексирования, а в некоторых случаях и единственным.

Основные выводы

1. Рассматривалась возможность проведения содержательной обработки документов КС в дополнение к традиционному предметному индексированию.
2. Предложено включать КС в ПОД в терминологии автора без нормализации.
3. КС следует рассматривать как дополнительное средство индексирования документа для:
   а) увеличения полноты индексирования
   б) представления побочных тем документа, не описанных другими языками индексирования
4. Обработка документов КС не является обязательной.
5. Процесс индексирования КС не требует больших дополнительных временных затрат на обработку документа.
6. Предложена методика выборки и форма представления КС применительно к ЭК.
7. КС могут быть особенно полезны для индексирования документов в локальных БД, в областных (например, отделы краеведения) и узкопрофильных библиотеках.

Литература

1. Гендина Н.И. Лингвинистическое обеспечение автоматизированных библиотечных систем. Алма-Ата, 1991. 221с.
2. Гринина Р.Ф., Соколов А.В. Сравнение предметных и дескрипторных информационно-поисковых систем//НТБ СССР. 1970. Вып. 3-4 (85-86). С.48-58.
3. Кругликова В.П. Предметизация произведений печати: Общ. методика. М. 1967. 173с.
4. Михайлов А.И., Черный А.И., Гиляревский Р.С. Основы информатики. М., 1968. 756 с.
5. Соколов А.В. Методические материалы по разработке информационно-поисковых тезаурусов. Учебн.-метод. пособие. Л.,1975.68с.
6. Aitchison, J., Gilchrist, A. Thesauri construction: A practical manual. 2-nd ed. 1987. 174 p.
7. Aluri R., Kemp D.A., Boll J.J. Subject analysis in online catalogs. 1991. 303p.
8. Linguistic approaches to text management: An appraisal of progress/Smeaton Alan F.//J. Doc. and Text Manag. 1994.V 2. №2. P. 67-80.
9. Needs for research in indexing/Milstead Jessical//J. Amer. Soc. Inf. Sci.1994.V45. №8. P.577-582.
10. Peters T.A., Kurth M. Controlled and uncontrolled vocabulary subject searching in an academic library online catalog// J.Information Technology and Libraries. September 1991. P.201-211.
11. Rowley J. The controlled versus natural indexing languages debate revisited: A perspective on information retrieval practice and research//J. Inf.Sci.1994. V.20. №2. P.108-119.
12. Squires S.J. Access to biomedical information: The Unified Medical Language System // Libr. Trends. 1993.V42. №1. P127-151.
13. Svenonius E. Precoordination or not?//Subject Indexing: Principles and Practices in the 90's. Munchen. New Providence. London. Paris. V.15.1995.P.231-255.
14. Un syteme d'indexation structuree a I' INIST. Bilan d'une etude prealable/Coret Annie, Menon Bruno, Schibler Daniele, Terrasse Christophe//Documentaliste. 1994.V31. №3. P.148-158.
15. Wanted: Fully automated indexing/Purcell royal//Libr. Software Rev.1991.V10. № 6.P.390-395