ОРГАНИЗАЦИЯ ИНФОРМАЦИОННОЙ РАБОТЫ

УДК 025.43:[001+62]

В. А. Быков, Е. Б. Дудин, М. М. Мельникова

Принципы составления списков основных ключевых слов по различным отраслям науки и техники

Рассматриваются принципы составления единых или совместимых списков основных ключевых слов, используемых для координатного индексирования документов при одноразовом реферировании — отбор лексики, выбор формы представления терминов (прямая и инверсная формы, единственное или множественное число), структура списков и т. п.

ОБЩИЕ ПОЛОЖЕНИЯ

При создании и эксплуатации информационно-поисковых систем особое значение имеет аналитико-синтетическая переработка документов — реферирование, редактирование, индексирование ключевыми словами и рубриками Рубрикатора и т. д. — для ввода результатов этой переработки в ЭВМ с целью генерации машиночитаемых баз данных (БД) и автоматизированной подготовки различных информационных изданий.

ВИНИТИ генерирует 242 машиночитаемые БД общим объемом ~1100 тыс. док/год и издает 249 выпусков Реферативного журнала, включающих ~1200 тыс. док/год [1].

Одним из основных аспектов аналитико-синтетической переработки документов является так называемое координатное индексирование, т. е. выражение основного смыслового содержания документа (реферата) в виде определенного набора ключевых и пояснительных слов [2]. Результатом координатного индексирования является поисковый образ документа, который служит как для поиска информации в БД ВИНИТИ, так и для генерации предметных указателей (печатных и/или машиночитаемых) к Реферативному журналу ВИНИТИ [3].

Принцип одноразового реферирования и индексирования документов и многократного использования результатов для генерации БД и информационных изданий предъявляет повышенные требования к качеству обработки первичной информации. Для этого необходимо хорошо разработанное лингвистическое обеспечение, в частности — списки основных ключевых слов по различным отраслям науки и техники, которые являются важнейшим инструментом координатного индексирования документов. В технологии одноразового реферирования списки ключевых слов по различным отраслям науки и техники должны быть едиными и/или совместимыми.

Для координатного индексирования документов из различных отраслей науки и техники в ВИНИТИ используются списки ключевых слов, которые зачастую составлялись на основе ручных картотек и/или с использованием машиночитаемых БД [4]. В настоящее время актуализация списков ключевых слов проводится только с использованием БД и единых правил представления лексики.

Списки подбираются путем "съема" ключевых слов из поисковых образов документов БД ВИНИТИ и последующей их экспертной оценки, с учетом частоты встречаемости термина. Для определения лексического поля списков следует выявлять не только основные ключевые слова данной отрасли науки или техники (или так называемое ядро терминологии), но и термины смежных отраслей, т. е. выявлять степень пересечения лексических полей различных отраслей науки и техники.

Это особенно важно в условиях одноразового реферирования и необходимости подготовки единых и/или совместимых списков по различным отраслям науки и техники и, в перспективе, тезауруса ВИНИТИ, так как в первую очередь необходимо согласовывать терминологию, встречающуюся в смежных отраслях. Кроме того, в тематических выпусках БД и/или РЖ, создаваемых на основе документов из различных областей знания, проблема единой терминологии стоит наиболее остро.

Списки предназначены для:

- использования при разработке и эксплуатации автоматизированных информационно-поисковых систем по различным отраслям науки и техники;
- унификации и стандартизации терминологии в конкретной области науки и техники;
- использования при координатном индексировании документов (составление поисковых образов) при генерации БД и предметных указателей к РЖ;
- поиска сведений в БД и предметных указателях (поскольку ключевые слова являются заголовками рубрик в предметных указателях).

В последнем случае единая форма представления ключевых слов особенно важна для систематизации сведений и удобства пользования предметным указателем.

Ключевыми словами (КС) считаются нормализованные слова и устойчивые словосочетания естественного языка, которые представляют собой термины или понятия конкретной отрасли науки и техники и несут в совокупности максимально полную и сжатую информацию о содержании документа (реферата). КС, используемые при индексировании, могут не содержаться в явном виде в тексте документа. При формировании печатных предметных указателей в качестве заголовков рубрик указателей используются только КС.

Ключевые слова, которыми индексируют реферат, должны отражать все основные аспекты его содержания, в том числе и те, которые являются "непрофильными" по отношению к тому разделу РЖ, где помещен индексируемый документ, но которые представляют интерес для специалистов смежных отраслей знания.

В качестве КС используются перечисленные ниже группы понятий, если в реферате содержится оригинальная информация об обозначаемом ими предмете:

1) объекты исследования (техника безопасности, производство и т. д.);

2) процессы и явления (отжиг, плавка, сварка, вибрация, шум и т. д.);

3) вещества и материалы (пластмассы, металлы и т. д.);

4) источники веществ, загрязняющих окружающую среду (транспорт, отрасли промышленности и т. д.);

5) характеристики процессов (кинетические параметры);

6) методы исследования, анализа (спектроскопия, дефектоскопия, моделирование математическое и т. д.);

7) приборы, аппараты, устройства, изделия (датчики, фильтры и т. д.);

8) законы, уравнения, функции (Больцмана уравнение, функция распределения);

9) общенаучные термины и прочие слова (обзоры, стандарты, конференции, персоналии и т. д.).

КС, представляющие собой общие понятия, контролируются по специальному словарю — "Списку основных ключевых слов", в котором устранены синонимия, полисемия и омонимия между терминами [5, 6].

В качестве КС могут использоваться как отдельные слова, так и словосочетания. Если в качестве КС используются словосочетания, то они включаются в "Список основных ключевых слов" как в прямой, так и в инверсной форме.

В инверсной форме на первом месте стоит слово, несущее основную смысловую нагрузку, а на втором — относящееся к нему прилагательное или существительное, например: *формы безопочные, формы гипсовые, формы керамические* и т. д.

Инверсный порядок слов целесообразен в тех случаях, когда его использование способствует группированию однородного по смыслу материала и повышает полноту поиска сведений в указателе.

В прямой форме существительное, несущее основную смысловую нагрузку, стоит на втором месте, а на первом — относящееся к нему прилагательное или существительное, например: *сверлильные станки, токарные станки, фрезерные станки.*

При использовании словосочетаний, включающих имя собственное, последнее должно стоять на первом месте, например: *Клейтона закон, Хорндала эффект.*

КС в "Списке основных ключевых слов" приведены к нормализованному виду, т. е. имеют форму именительного падежа единственного или множественного числа. **В единственном числе** используются неисчисляемые термины, обозначающие: процессы, научные дисциплины, отрасли промышленности, названия конкретных веществ или материалов, характеристики процессов и т. п.

Например: *программирование, теоретическая механика, алюминий, динамика, инфляция* и т. д.

Во **множественном числе** используются исчисляемые термины, обозначающие: приборы, устройства; групповые названия материалов, веществ и т. п.

Например: *фильтры, станки, металлы* и т. д.

В остальных случаях КС могут иметь форму как единственного, так и множественного числа, в зависимости от того, выражает термин широкое (родовое) или узкое (видовое) понятие. Примеры: *Оптическая активность, Нееля температура, Зеемана эффект,* но: *Оптические свойства, Критические постоянные, Гальваномагнитные явления.*

Этими же принципами для выбора формы единственного или множественного числа следует руководствоваться при индексировании документов ключевыми словами, которые не вошли в "Список основных ключевых слов", т. е. если предметное понятие не имеет близкого эквивалента в "Списке основных ключевых слов", его следует описать в терминах, используемых в реферате и привести к нормализованному виду (именительный падеж, единственное или множественное число).

В качестве КС используются только общепринятые аббревиатуры, например, ЭВМ, ООН и т. д. Использование сокращений в КС не допускается.

В ряде случаев в список не включают многие термины, используемые в качестве ключевых слов при индексировании. К ним могут относиться: наименования географических объектов, например *Баренцево море, Волга река, Сингапур* и т. д., наименования конкретных химических соединений, названия конкретных растений и организмов.

Это вызвано тем, что число подобных терминов очень велико, например, количество химических соединений достигает 18 млн. В таких случаях в предисловии к спискам приводят правила использования этих терминов. Например, наименования географических объектов пишутся с имени собственного с указанием характера объекта: *Тихий океан, Темерник город, Темерник река* и т. д.

Что касается наименований химических соединений, то в списки, как правило, включают наиболее часто употребляемые соединения, а классы соединений включают в отдельные списки. Исключение составляют некоторые термины, имеющие важное значение в науке и технике или отражающие очень общие и обширные понятия, например: *Вода, Смолы.*

Список основных ключевых слов имеет единую упорядоченную формализованную логическую структуру, в нем устранены синонимия, полисемия и омонимия между отдельными терминами, все слова приведены в стандартной словарной форме, т. е. в именительном падеже единственного или множественного числа, и в некоторых случаях между ними установлены смысловые связи.

ФОРМАЛЬНАЯ СТРУКТУРА СПИСКА

Все ключевые слова, входящие в словарную часть, печатаются с левой стороны колонки прописными буквами и расположены в алфавитном

порядке, например: *МЕЛИОРАЦИЯ, МЕМБРАННЫЕ ТЕХНОЛОГИИ.*

Встречаются случаи, когда после ключевого слова и отсылок к нему помещается в алфавитном порядке набор слов, напечатанных строчными буквами, которые являются примером типичных подзаголовков рубрик указателя или пояснительных (неключевых) слов и часто представляют собой элементы классификации основного термина, например: *НАЦИОНАЛЬНЫЕ ПАРКИ (городские, природно-исторические, природные* и т. д.).

Этот список подзаголовков или пояснительных слов раскрывает структуру и содержание рубрик конкретного указателя.

ИСПОЛЬЗОВАНИЕ ОТСЫЛОК "СМ."

Отсылки *см.* используются в следующих случаях:

- для устранения полной синонимии, т. е. для унификации терминологии, например: *сосняки см. СОСНОВЫЕ ЛЕСА*;
- в случае частичной синонимии многозначных терминов, или когда термины очень близки по значению, например: *упаковка см. УПАКОВОЧНЫЕ МАТЕРИАЛЫ;*
- в случае индексирования какого-либо аспекта документа вышестоящим понятием, т. е. при замене видового понятия родовым, например: *угледобыча см. УГЛЕДОБЫВАЮЩАЯ ПРОМЫШЛЕННОСТЬ*. Здесь отсылка *см.* означает, что в предметных указателях все сведения об угледобыче собираются в рубрике *УГЛЕДОБЫВАЮЩАЯ ПРОМЫШЛЕННОСТЬ*.

ИСПОЛЬЗОВАНИЕ ОТСЫЛОК "СМ. ТАКЖЕ"

Отсылки "*см. также*" выражают логические связи между ключевыми словами, например "вид — род", "часть — целое", "причина — следствие" и т. д. и используются для раскрытия структуры указателей в следующих случаях:

- для отражения разбиения крупных рубрик на более мелкие, например: *СТОЧНЫЕ ВОДЫ (см. также ГОРОДСКИЕ СТОЧНЫЕ ВОДЫ, ПРОМЫШЛЕННЫЕ СТОЧНЫЕ ВОДЫ* и т. д.)
- для связывания родственных рубрик в указателе.

В этом случае отсылка *"см. также"* означает, что для увеличения полноты и точности поиска нужной информации читателю полезно просмотреть соответствующие рубрики, связанные перекрестными ссылками. Например: *ОКРУЖАЮЩАЯ СРЕДА (см. также БИОСФЕРА, ОБЪЕКТЫ ОКРУЖАЮЩЕЙ СРЕДЫ, ЭКОСИСТЕМА* и т. д.)

ИНВЕРСИЯ СЛОВОСОЧЕТАНИЙ

Термины, состоящие из двух и более слов, в списке представлены в прямой или в инверсной форме. Это сделано с целью выделения слова, несущего наибольшую смысловую нагрузку, а также с целью унификации терминологии. Если наиболее существенную часть словосочетания выделить трудно, то в списке приводятся обе формы — прямая и инверсная с указанием, какая форма принимается за основную, например: активированный уголь *см. УГОЛЬ АКТИВИРОВАННЫЙ*

ЗАКЛЮЧЕНИЕ

Применение списков ключевых слов в технологии ВИНИТИ имеет большие перспективы в возможной системе автоматизированного контроля и корректуры нормализованной лексики поисковых образов документов в базах данных. Такая система, логически необходимая в технологии однократного реферирования, может обеспечить:

- улучшение качества информационных продуктов (баз данных и предметных указателей) и, как следствие, повышение их конкурентоспособности;
- уменьшение трудозатрат при подготовке баз данных и предметных указателей;
- усовершенствование структуры и технологии подготовки предметных указателей.

На выходе такой системы могут быть получены:

- актуализированные массивы нормализованной лексики в электронной или печатной форме;
- откорректированные поисковые образы документов в базах данных с нормализованной лексикой;
- принт-файлы и оригинал-макеты номерных и годовых предметных указателей.

Списки ключевых слов могут стать основой толковых словарей по различным отраслям науки и техники, а также основой двуязычных словарей, используемых при автоматизированном или обычном переводе. При систематическом контроле нормализованной лексики по частоте появления новых терминов можно судить об изменении направлений исследования данной отрасли науки или техники, ее точках роста, изменении терминологического поля и т. д. Эти статистические данные могут быть использованы при принятии решения об актуализации лингвистического обеспечения системы подготовки информационных продуктов, рубрикатора отрасли науки или техники и т. п., о согласовании терминологии в смежных отраслях знания, при решении вопросов генерации проблемных фрагментов баз данных и/или подготовки выпусков реферативных журналов и т. д.

СПИСОК ЛИТЕРАТУРЫ

1. Проспект информационных изданий ВИНИТИ 2002–2003 гг.— М., 2002.— С. 96.

2. НТП ВИНИТИ 9–96 "Координатное индексирование. Основные положения".— М.: ВИНИТИ, 1996.

3. НТП ВИНИТИ 8–96 "Указатели к Реферативному журналу ВИНИТИ. Основные положения".— М.: ВИНИТИ, 1996.

4. Бондарь В. В., Мельникова М. М., Гончарук Г. П., Ибрагимова М. Б., Ренард Т. Л., Соковикова Н. К., Цветкова И. Д. Принципы отбора лексики для словарей и лексических пособий по химии и химической технологии // НТИ. Сер. 1.— 1984.— № 12.— С. 24–28.

5. Список основных ключевых слов по химии и химической технологии (пятый вариант).— М.: ВИНИТИ, 1998.

6. Список основных ключевых слов для координатного индексирования документов по проблемам охраны окружающей среды и экологии.— М.: НТП ВИНИТИ, 2002.

Материал поступил в редакцию 22.07.2002.