Анализ соответствия специальности авторефератов диссертаций

Леонова Ю.В.

Введение

Научная информация представляет собой особый класс информации, которая в отличие от понятия информации в широком смысле слова всегда в той или иной степени организованна, т.е. обобщена, классифицирована, свернута, другими словами – обработана. Под научно информационный деятельностью понимается «самостоятельная часть научного труда» выделившаяся в ходе его общественного разделения и имеющая свои задачи обеспечения заинтересованных специалистов необходимыми им сведениями о результатах научных исследований и технических разработок, о производственном опыте. Объектом научной информации в этом плане следует считать как исходные данные (первичная информация), так и обобщенные материалы, результаты обобщения, анализы, классификации, свертывание первичной информации.

Особый интерес для ученых представляют непубликуемые научные документы (или так называемая «серая литература»), которыми являются диссертации. Эти документы содержат большое количество ценной информации, значительная часть которой не попадает на страницы научных изданий.

Диссертацией называется научное исследование, представленное на соискание ученой степени, *автореферат*ом - изложение основных положений диссертации, составленное автором для предварительного ознакомления с ними научной общественности.

В нашей стране диссертации как таковые не публикуются, но подвергаются строгой библиографической регистрации. Хотя они существуют в единичных экземплярах, но изложенные в них идеи и факты считаются официально введенными в научный оборот. Авторефераты диссертаций публикуются ограниченным тиражом (100-150 экземпляров). Они помечаются грифом «На правах рукописи», однако обладают всеми правами произведения печати.

В базе данных Публикаций СО РАН содержится более 90 тысяч описаний диссертаций и авторефератов.

- ✓ Годы: 1983-2011
- ✓ Схема данных: название, автор, год, место защиты, дата защиты, ученый совет, специальность, ученая степень, количество страниц, организация выполнения работы, научные руководители, оппоненты, ведущая организация, оглавление, введение, заключение, список литературы, текст автореферата.

Информационный поиск и классификация

При решении задач информационного поиска в документах научно-технического и образовательного характера важной задачей является отнесение документа к одной из нескольких категорий на основании семантического содержания документа, называемой задачей классификации. Классификация документов позволяет установить специфику каждого вида и типа документа, помогает ориентироваться во всем многообразии документной информации. В

процессе классификации содержание документов должно анализироваться как с позиций того, какие идеи и факты заложены в него автором, так и с позиций научных и практических интересов большинства его потенциальных читателей (Если не иметь в виду узкоспециальных интересов, то обе точки зрения чаще всего совпадают). Средством классификации понятий в определенной области является тезаурус, который содержит:

1. структурированную систему понятий с определением иерархических и ассоциативных отношений между понятиями;
2. список терминов, определяющих каждое из понятий; все термины, определяющие одно и то же понятие, называются синонимами; Сведение синонимов вместе реализует терминологический контроль, обеспечивающий возможность выражения одного и того же понятия разными способами.

Тезаурус содержит набор дескрипторов для индексирования и поиска документов.

Современный этап развития систем информационного поиска ведет начало от работ К. Мауэрса, предложившего описывать содержание документов простым перечислением дескрипторов – терминов, употребляющихся в самом документе и тем самым определяющих его содержание в пределах терминологии данной предметной области. Предполагалось, что перечень дескрипторов, существенных для описания документов определенной предметной области, не слишком велик, что его можно свести в словарь, в котором будут заданы отношения между дескрипторами наподобие соотношения иерархии классов понятий, и использовать этот ограниченный словарь, называемый информационно-поисковым тезаурусом, как систему фасетного индексирования документов. Специфика некоторых тематических областей, например, сферы науки и образования, заключается, в частности, в том, что ее терминология не имеет узкоотраслевого характера, а включает в себя наряду со специальной научной и педагогической лексикой, также все термины тех отраслей знания, по которым ведется исследование и обучение. Поэтому наряду с узкоотраслевыми тезаурусами должны использоваться универсальные тезаурусы научной и технической лексики. Очевидно, что использовать неопределенный набор тезаурусов в качестве единой классификационной системы невозможно.

В современном подходе состоит в использовании тезаурусов не для индексирования документов, а для определения релевантности документа поисковому запросу. При этом классификационными признаками документов служат сами слова, которые употреблены в документе (ключевые слова), а критерий соответствия документа запросу определяется на основе семантической информации по возможности обо всех ключевых словах данной специальности. Каждый пользователь , заинтересованный в работе по своему направлению (специальности), должен дополнять встроенный в систему общетехнический словарь только своим специфическим словарем.

Классификация фактов

Рассмотрим задачу классификации диссертаций и авторефератов по специальностям ВАК. Поскольку в ВАК нет формальных критериев, по которым можно определить соответствие диссертации определенной специальности, то построить тезаурус ключевых слов для классификации диссертаций по всем специальностям затруднительно. В ВАК существуют только лексические критерии соответствия формулировок документов и результатов исследования. При этом формулировки одних и тех же результатов с небольшими отличиями могут соответствовать разным специальностям. Кроме того, в разных ученых советах предъявляются разные требования к диссертациям по одной и той же специальности.

Одним из способов классификации диссертаций может быть классификация фактов, содержащихся в диссертациях в соответствии со специальностями ВАК. Для выделения фактов используется принцип координатного индексирования – выражения основного смыслового

содержания (предмета) документа в виде определенной совокупности ключевых слов. Различные реализации метода индексирования по ключевым словам предполагают выделение информативных слов документа и группы слов, например, с наибольшими значениями частот включают в список ключевых слов документа. Вместе с тем содержание документа лучше описывается не отдельными словами, а понятиями и терминами, представляющими собой совокупность слов или словосочетания. В информационных системах широко используются понятия, выраженные именными словосочетаниями. В именных словосочетаниях главным словом (основным носителем смысла) является, как правило, первое слева существительное, а остальные слова служат для уточнения значения главного слова.