Информационно-поисковые языки в электронной среде: новые возможности

Скарук Г А.

ГПНТБ СО РАН (Новосибирск)

Эффективность использования информационных ресурсов библиотек и качество информационного сервиса как на локальном уровне, так и в режиме сети во многом зависит от качества и эффективности использования информационно-поисковых языков (ИПЯ) основных библиотечных банков данных – электронных каталогов (ЭК).

Согласно общим принципам проектирования автоматизированных информационно-поисковых систем, процедура формирования лингвистического обеспечения (ЛО) должна осуществляться на основе анализа следующих факторов: функционально-целевого назначения системы, характера поступающих в нее запросов, способности различных типов ИПЯ обеспечить их реализацию. Однако на начальных стадиях создания российских ЭК углубленных исследований названных факторов проведено не было. Возникающие проблемы осознавались и решались в основном на эмпирическом уровне. В условиях же активного функционирования библиотечных корпоративных сетей различного уровня, что выводит на первый план вопрос о национальной политике в области поддержания и использования лингвистических средств, проблемы оптимизации ЛО приобрели более четкие постановочные аспекты.

Основная предпосылка изменения свойств ИПЯ в электронной среде, – изменение принципов организации поискового массива и реализации поисковых процедур. К наиболее важным следствиям этого можно отнести: множественность «точек доступа» к библиографическим записям, возможность логического сочетания поисковых признаков; возможность использования в качестве поисковых реквизитов как полных индексов, так и отдельных элементов комбинированных индексов предкоординатных ИПЯ; возможность применения критериев релевантности на неполное соответствие; возможность усечения индексов, отделение поискового словаря от поискового массива, возможность упорядочить информационный массив в зависимости от поисковой задачи по любому заданному признаку. В связи с этим изменяются как свойства ИПЯ, так и требования ним и способы реализации этих требований. Но последствия «компьютерной революции», особенно в сфере использования предкоординатных ИПЯ традиционных каталогов, на наш взгляд, и сегодня недооцениваются специалистами, поэтому значительная часть преимуществ автоматизированного поиска по сей день остаются гипотетическими, методически и программно не обеспеченными.

Эксперимент по сопоставлению функциональных возможностей Государственного рубрикатора научно-технической информации (ГРНТИ), Библиотечно-библиографической классификации (ББК), языка предметных рубрик (ЯПР) и языка ключевых слов из заглавий

документов (ЯКС), проведенный в ГПНТБ СО РАН, позволил уточнить причины и следствия изменения характеристик ИПЯ, определить содержание необходимых преобразований, наметить меры по их оптимальному комплексному использованию.

Процедура эксперимента. Реальные читательские запросы для экспериментального тестирования были отобраны случайным образом. В выборке представлено в совокупности 100 запросов: по 10 запросов следующей отраслевой принадлежности:

социально–гуманитарные науки – философия и психология; социология; история; экономика; право;

естественные и технические науки: – электро-, радиотехника и электроника; химическая технология; науки о Земле (геология, география); биология; медицина.

Самостоятельное поисковое предписание составлялось и реализовывалось с использованием каждого из участвующих в эксперименте поисковых языков. Далее определялась общая (суммарная) выдача по запросу путем логического объединения выдач, полученных в результате поиска на отдельных ИПЯ. Релевантность выданных документов определялась по двубальной шкале: документ признавался либо релевантным запросу, либо нерелевантным.

Процедура анализа данных. На первом этапе анализа проводилась диагностика запросов. Для каждого запроса определялись отраслевая принадлежность и детальная категориальная структура. По результатам поиска по каждому запросу на каждом из участвующих в эксперименте ИПЯ вычислялись коэффициенты точности и относительной полноты. Такой подход в данном случае представляется вполне оправданным, так цель эксперимента состоит не в определении абсолютных показателей, а в их сравнении. Для расчета коэффициентов были использованы следующие формулы:

Относительная полнота поиска по запросу определялась по формуле (1):

$$R_q = \frac{r_q}{r_s} \qquad (1)$$

где:

R_q – относительная полнота поиска по запросу **q**;

r_q – число релевантных документов в выдаче по запросу на отдельном ИПЯ;

r_s – число релевантных документов в суммарной выдаче по запросу.

Точность поиска вычислялась по формуле (2):

$$P_q = \frac{r_q}{s_q} \qquad (2)$$

где: P_q – полнота поиска по запросу **q;**

$\mathbf{r_q}$ – число релевантных документов в выдаче по запросу на отдельном ИПЯ;

$\mathbf{s_q}$ – общий объем выдачи по запросу на отдельном ИПЯ.

Следующий этап исследования предполагал выявление причин поисковых неудач в зависимости от типа запроса. Выяснялись причины потерь релевантных и выдачи нерелевантных документов в результате поиска на отдельных ИПЯ. За эталон принималась суммарная выдача по запросу. При этом выделялись поисковые неудачи:

- обусловленные возможностями лексики и грамматики собственно поискового языка;
- связанные с особенностями индексирования документа на соответствующем ИПЯ;
- порожденные неточностями, допущенными в процессе составления поискового предписания (индексирования запроса).

Для двух последних категорий неудачных поисков особо оценивалось, были ли они обусловлены требованиями методики индексирования или стали следствием субъективного подхода индексатора. В ходе анализа качества семантической обработки сравнивались показатели полноты и глубины индексирования каждого документа на различных ИПЯ.

Итоги и выводы. В результате эксперимента подтверждена целесообразность применения в ЭК комплекса разнообразных по структуре, свойствам, лексическому наполнению лингвистических средств.

Таблица 1. Показатели эффективности поиска на различных ИПЯ в ЭК ГПНТБ СО РАН

ИПЯ	БД книг и сборников		БД авторефератов диссертаций	
Коэффициенты эффективности поиска	Полнота	Точность	Полнота	Точность
ГРНТИ	68,6	25,2	44,8	29.8
ББК	74,3	73,9	64,0	69,7
ЯПР	67,5	90,1	44,8	76,5
ЯКС	54,8	79,1	49,2	64,6

Как следует из таблицы 1, в целом по экспериментальной совокупности наибольшая полнота поиска достигнута с использованием ИПЯ ББК, наибольшая точность – в результате поиска по предметным рубрикам. Язык ключевых слов занял среднюю позицию по эффективности поиска, значения коэффициентов полноты и точности в среднем приближались к 50% или несколько превышали этот показатель. При этом необходимо учитывать, что среднее количество ЛЕ в ПП на этом языке составило 9,4. Эта цифра значительно превышает аналогичные показатели для ББК и ЯПР. Очевидно, что достижение полученных в эксперименте результатов будет сопряжено для конечных пользователей ЭК со значительно большими, чем при использовании контролируемых ИПЯ, трудностями.

Сравнение свойств классификационных и предметизационных ИПЯ (таблица 2) обеспечило возможность уточнить их функции по реализации стратегий, ориентированных на преимущественное достижение тех или иных параметров выдачи, и организации помощи пользователям ЭК в процессах формулирования и корректировки ПП.

Таблица 2. Преимущества поиска на предкоординатных ИПЯ

ББК	ЯПР
наибольшая полнота поиска по совокупности запросов	наибольшая точность поиска по совокупности запросов
для запросов по естественным и техническим наукам	для запросов по социально-гуманитарным наукам
для сложных понятий	для простых понятий

Эксперимент показал, что в результате автоматизации поиска сокращаются различия между этими языками в части локализации поисковых признаков (предмет или отрасль), но принципиальные отличные возможности возникают при поиске с усечением. Таким образом, ведущие различия между ИПЯ определяются сегодня не по способу пред- или посткоординации, а по принципам плана выражения и структурной иерархии.

Причины информационного шума (таблица 3) и потерь информации (таблица 4) распределились следующим образом:

Итак, в качестве базовых компонентов ЛО ЭК можно предложить комбинационную иерархическую классификацию и язык предметных рубрик по соображениям различия механизмов помощи в формулировании запросов и отборе релевантной информации, в разной степени востребованных в конкретных поисковых ситуациях. Обнаруженные в результате экспериментального тестирования ИПЯ различия в их поисковой эффективности при реализации различных типов запросов также подтверждают этот тезис.

Таблица 3. Распределение средних значений причин информационного шума при поиске на различных ИПЯ (в %)

БД	Причины ИПЯ	Обусловленные свойствами ИПЯ					Обусловленные ка индексирования до	
		Специфичность ЛЕ	Ложн. координация ЛЕ	Полисемия	Неинформативность	Другие	Полнота	
							Методические	Су
БД книг	ГР	93	–	–	–	–	–	-
	ББК	91	–	–	–	–	–	-
	ЯПР	14	17	–	–	–	–	-
	ЯКС	–	11	4	9	5	–	-
БД авторефератов	ГР	90	–	–	–	–	–	-
	ББК	85	–	–	–	–	–	-
	ЯПР	12	6	3	–	–	6	-
	ЯКС	–	11	1	1	–	–	-

Таблица 4. Распределение средних значений причин потерь информации при поиске на различных ИПЯ (в %)

		Обусловленные свойствами ИПЯ				Обусловленные качеством индексирования документа		
		Невозможность выразить запрос средствами ИПЯ	Специфичность ЛЕ	Неинформативность	Другие	Полнота		Гл
						Методические	Субъективные	Мето

БД книг	ГР	–	–	–	–	1	46	8
	ББК	–	–	–	–	4	21	22
	ЯПР	–	–	1	6	2	24	13
	ЯКС	–	–	20	9	1	4	–
БД авторефератов	ГР	1	–	–	–	–	28	2
	ББК	1	–	–	1	10	26	7
	ЯПР	–	1	–	3	–	35	2
	ЯКС	–	–	–	–	–	28	2

Наиболее эффективным при использовании в ЭК языков традиционных библиотечных каталогов представляется сочетание принципов предкоординации и посткоординации, то есть применения в зависимости от параметров поисковой ситуации как полных индексов ИПЯ, так и их элементов в качестве самостоятельных поисковых реквизитов с усечением каждого из них. Но если для языка предметных рубрик такие возможности уже заложены в существующих АБИС, то для иерархических классификаций проблема практически не решена.

Анализ неудачных поисков продемонстрировал, что применение традиционных методик индексирования документов, разработанных в свое время для карточных каталогов, ограничивает поисковые возможности ИПЯ в электронной среде. Необходим пересмотр методик в направлении расширения содержания поискового образа документа.

Принципы разыскания документов различны для запросов различной категориальной структуры, видов затребованных документов. Тем самым доказывается целесообразность детальной многоаспектной диагностики запросов конечных пользователей ЭК.

Основное требование к комплексу ИПЯ еще на начальных этапах становления концепции ЛО АБИС формулировалось как целостность, заключающаяся в максимально возможных взаимодействии и связанности все ИПЯ и сохранении такой связанности при их использовании (51).

Проведенное исследование позволило оценить возможность приложения и степень проявления этого требования к различным объектам и этапам поискового взаимодействия и на этой основе сформулировать некоторые принципы и рекомендации по комплексному использованию лингвистических средств тематического поиска в электронном каталоге.

Результаты эксперимента показали, что *принцип взаимодополнения* ИПЯ необходимо соблюдать:

- при определении состава и параметров ЛО;
- при составлении поискового предписания (с ограничениями, так как в поисковом образе запроса возможно не только комплексное, но и самостоятельное применение отдельных языковых средств);
- при составлении поискового образа запроса.

В последнем случае целесообразно применять *принцип дифференциации функций отдельных ИПЯ* при поиске в зависимости от его функциональных свойств и возможностей, необходимых и достаточных в конкретной поисковой ситуации.

С этой целью поисковая ситуация описывается на основании *принципа комплексной*

диагностики, согласно которому процедуру диагностирования должна включать диагностику как самого запроса, так и предъявившего его пользователя. В обоих случаях необходима оценка комплекса характеристик, а именно:

- при диагностировании запроса – его отраслевой принадлежности, состава, содержания и объема включенных в него семантических категорий, вида искомых документов;
- при диагностировании пользователя – цели поиска, уровня знания темы поиска, наличия поискового опыта.

Принцип комплексной оценки свойств и функциональных возможностей отдельных ИПЯ подразумевает рассмотрение возможностей ИПЯ по обеспечению как эффективности, так и комфортности поиска. С этих позиций оценке подлежат:

- *свойства самого ИПЯ*, в том числе:

- семантическая сила ИПЯ (состав и специфичность ЛЕ, состав и функции грамматических средств, наличие механизмов многоаспектного представления содержания документа / запроса средствами ИПЯ);
- способность обеспечить результаты требуемого уровня качества для различных типов поисковых запросов;

- *состояние и содержание методик индексирования* на различных ИПЯ, в том числе:

- степень адаптации к ситуации автоматизированного поиска;
- степень ориентации на потребности пользователя ЭК;
- предусмотренные ограничения в полноте и точности описания содержания документа/запроса;

- *возможности манипулирования ИПЯ неподготовленными пользователями ЭК*, в том числе

- объем и структурная сложность лексики и грамматики поискового языка;
- наличие механизмов экспликации смысла ЛЕ;
- наличие механизмов экспликации структуры и состава ИПЯ;
- степень соответствия структуры ЛЕ и нормативных словарей ИПЯ естественным речемыслительным структурам;
- наличие механизмов уточнения темы запроса;

- *состав и содержание типичных ошибок при индексировании документов и запросов;*

- *возможности представления и коррекции свойств ИПЯ программными средствами*, в том числе:

- возможности поиска с усечением;
- возможности использования в режимах пред– и посткоординации;
- возможности доступа к авторитетным файлам ИПЯ.

Как составляющую принципа дифференцированного использования можно

рассматривать *принцип дифференцированного доступа*, в соответствии с которым экспликация для пользователей состава и структуры ИПЯ должна проводиться по результатам диагностики поисковой ситуации.

Из принципа комплексной оценки закономерно вытекает *принцип комплексного преобразования* ЛО ЭК, согласно которому в случае преобразования одного из компонентов ЛО следует оценивать его виляние на все перечисленные при формулировании принципа комплексной оценки характеристики каждого из элементов ЛО, и, при необходимости, осуществлять согласованное преобразование всего комплекса ИПЯ.

Оценку результатов применения названных принципов целесообразно проводить на основе *принципа комплексного контроля качества индексирования*. Анализ ошибок, допущенных при индексировании документов, позволяет говорить о необходимости в процессе контроля качества результатов семантической обработки соотносить состав и структуру ПОД на различных ИПЯ. Эталоном в этом случае выступает ПОД, в результате декодирования которого получено наиболее полное описание содержания документа. Важно оценить: соотношение смыслов и объемов понятий, полученных в результате декодирования, соотношение количества ЛЕ, использованных для описания основного предмета и аспектов содержания документа. Далее оценивается, имеется ли возможность отразить выявленные пробелы в описании содержания средствами каждого поискового языка и целесообразно ли такое дополнение в ситуации комплексного использования ИПЯ. Правила взаимодополнения индексов на различных ИПЯ должны быть четко регламентированы в соответствующих нормативно–методических документах и учтены затем при составлении поискового предписания.

Из последнего положения следует *принцип координации методик индексирования на отдельных ИПЯ* с учетом как специфических функций ИПЯ в составе комплекса, так и задач их сопряжения.

По нашему мнению соблюдение вышеназванных принципов будет способствовать эффективному и экономичному использованию комплекса ИПЯ в целом и отдельных поисковых языков в его составе.