

Принятые сокращения и определение понятий.

НТИ – научно-техническая информация, включает патенты, непатентную литературу и иные нетекстовые данные, зарегистрированные, полученные, посвященные или связанные с научными или техническими проблемами.

НПЛ – непатентная литература включает любые (кроме патентов) затрагивающие научные или технические вопросы тексты, а также нетекстовые материалы с текстовым описанием или метаданными.

Метаданные – информация об информации, например, библиографическая запись (заголовок, автор, даты, коды классификаторов...), реферат, описание материала (текста, видеозаписи, звукозаписи, изображения, топологии, структуры...).

Поиск в тексте – поиск в базе данных записей, содержащих лексику, удовлетворяющую условиям поискового запроса. Если поисковый запрос не содержит элементов синтаксиса языка запросов конкретной базы данных, такой запрос можно назвать интуитивным. Пример интуитивного запроса (выделен курсивом с подчеркиванием): *cutting of carbon nanotubes*. В случае использования в запросе элементов синтаксиса языка запросов конкретной базы данных, запрос можно назвать сложным. Только сложный запрос может обеспечить качественный поиск в текстах. Пример сложного запроса для поиска англоязычных текстов в базе данных <https://patents.google.com>: *"cutting carbon nanotubes" OR "cutting of carbon nanotubes" OR "carbon nanotubes cutting"*.

Классификационный поиск – выборка в базы данных записей, отмеченных подходящими пользователю индексами / кодами классификатора. Содержательная разметка НТИ с использованием классификаторов именуется рубрицированием (отнесение к рубрике классификатора) или кодификацией (приписывание кода классификатора). Рубрицирование выполняется вручную, с использованием средств поиска в тексте или иных способов, или в смешанном режиме.

Точность поиска – субъективная оценка приемлемости для пользователя результатов поиска. Точность поиска - доля полезных документов среди найденных в конкретной базе данных. Низкая точность поиска фатальна: по опросам российских пользователей правовых баз данных после трех неудачных попыток они прекращают использовать конкретный способ поиска – времени жалко. Точность поиска можно оценить для любого способа поиска. Объективная оценка точности поиска требует привлечения группы специалистов с последующим усреднением их экспертных оценок. Ограничение: оценивающий точность поиска специалист должен уметь доказать несоответствие найденного поставленной задаче, что, к сожалению, случается нечасто и зависит как от уровня его формально-логической и предметной подготовки, так и от его добросовестности и мотивированности. Дополнительные сложности при оценке точности поиска создает как неполнота и однозначность постановки задачи, так и субъективность понимания этой задачи конкретным специалистом.

Полнота поиска – доля найденных полезных материалов из общего подмножества полезных материалов в конкретной базе данных. С учетом колоссального числа записей в конкретных базах НТИ, реальна только сравнительная оценка полноты поиска для, например, конкурирующих подходов к нахождению подходящих материалов или запросов для поиска в текстах. Чем больше найдено полезных материалов конкретным способом, тем лучше. Таким образом, на тестовой коллекции текстов, например, совокупности всех всемирных патентов, можно сравнивать полноту найденного семантическим поиском (искусственный интеллект / нейронные сети), поиском в текстах по конкретным поисковым запросам и классификационным поиском.

Качество поиска – это совокупность полноты и точности найденного. С учетом вероятностной модели естественного языка В.Налимова достижение 100% качества поиска информации на естественном языке невозможно. Чем выше точность поиска в тексте для заданной совокупности искомой лексики, тем ниже его полнота. Поэтому в производстве запросов для качественного поиска в тексте может применяться следующая норма: при максимальной полноте результатов поиска с учетом ограничений языка запросов конкретной базы данных, допустимого числа и предельной длины и сложности отдельного поискового запроса точность поиска должна быть не хуже 90%. При такой норме возможно без существенного понижения качества логическое пересечение множеств результатов поиска по разным темам и автоматическое рубрицирование потоков и массивов текстов.

ВОИС – Всемирная организация интеллектуальной собственности <https://www.wipo.int/> регистрирует всемирные патенты.

МПК – Международная патентная классификация ведется ВОИС и включает 74 тыс. рубрик <https://ipcpub.wipo.int/>

ЕПВ – Европейское патентное ведомство <https://www.epo.org/> регистрирует европейские патенты.

КПК – Кооперативная патентная классификация развивается ЕПВ и патентным ведомством США на основе МПК, включает 260 тыс. рубрик <https://worldwide.espacenet.com/patent/cpc-browser#>

Индекс МПК / КПК – рубрика **нижнего уровня** конкретного классификатора, не имеющая подрубрик / дочерних рубрик (кроме H10K50/00).

ПБД – патентная база данных:

ВОИС ПБД международной организации с 110 млн. патентных документов <https://patentscope.wipo.int/>

ЕПВ ПБД международной организации с 140 млн. патентных документов <https://worldwide.espacenet.com/>

БНТИ – база НТИ:

Гугл-Патенты коммерческая БНТИ б/п доступа с 120 млн. патентов + 400 млн. НПЛ <https://patents.google.com>

ВВМ - число найденных 25.05.2023 в ПБД ВОИС всемирных патентов с индексом МПК А61К9/51 по запросу типа: IC:(C01B32/176) Offices PCT/WO

ВЕК - число найденных 24.05.2023 в ПБД ЕПВ всемирных патентов с индексом МПК или КПК по запросу типа: pn=WO src=C01B32/176

ВМ - число найденных 26.05.2023 в ПБД ВОИС патентов с индексом МПК по запросу типа: IC:(C01B32/176)

ЕК - число найденных 24.05.2023 в ПБД ЕПВ патентов с индексом КПК по запросу типа: src=C01B32/176

ЕП - число найденных 24.05.2023 в ПБД ЕПВ патентов по рубрике КПК запросом для поиска в тексте (см. далее)

ГП - число найденных 28.05.2023 в БНТИ Гугл патентов по рубрике КПК запросом для поиска в тексте (см. далее)

ГН - число найденных 28.05.2023 в БНТИ Гугл единиц НПЛ по рубрике КПК запросом для поиска в тексте (см. далее)

ВЕК-ЕП – число кодифицированных по рубрике МПК / КПК всемирных патентов ПБД ЕПВ, не найденных запросом для поиска в тексте по рубрике МПК / КПК (см. далее тестовый сложный запрос № 2) и по результату изучения текста патента не соответствующих этой рубрике. Иными словами, это число характеризует уровень информационного шума классификационного поиска и позволяет оценить точность рубрицирования всемирных патентов в базе ЕПВ.

Таблица 1. Материалы для оценки качества кодификации НТИ с использованием кодов международных патентных классификаций из области «Нанотехнологии»

	Индекс КПК	Индекс МПК	ВВМ	ВЕК	ВЕК/ВМ	ВМ	ЕК	ЕП	ЕК/ЕП	ГП	ЕК/ГП	ГН	ЕК/(ГП+ГН)	ВЕК-ЕП	(ВЕК-ЕП)/ВЕК
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1.	B82B1/006	-	-	2	-	-	10								
2.	B82B1/007	-	-	3	-	-	4								
3.	B82B3/008 5	-	-	3	-	-	33								
4.	B82Y5/00	B82Y5/00	1085	5208	480%	10304	11037								
5.	B82Y10/00	B82Y10/00	672	7874	1172%	7330	25424								
6.	B82Y30/00	B82Y30/00	2865	16030	559%	54102	54343								
7.	B82Y35/00	B82Y35/00	87	622	714%	1153	1976								
8.	B82Y40/00	B82Y40/00	2545	9982	392%	56810	48927								
9.	C01B32/176	C01B32/176	8	29	363%	87	74	5661	1,3%	20176	0,36%	26335	0,16%	12	59%
10.	G01Q70/08	G01Q70/08	161	17	11%	1607	71								
11.	G02F1/01741	-	-	3	-	-	12								
12.	H01L29/732	H01L29/732	206	189	92%	5516	1036								
13.	H10K50/00	H10K50/00	0	6404	∞	0	10436								
14.	H10K50/115	H10K50/115	0	843	∞	0	3036								

Разъяснения к таблице 1:

- Ячейки заполнялись данными, которые можно проверить, выполнив тестовые запросы к базам данных (см. далее) или проведя расчеты.
- Экстремальные значения ячеек (не более 3) выделены жирным шрифтом, а самое крайнее значение выделено цветом.
- Ячейки не заполнялись, если данные не были получены.

Комментарий к таблице 1:

- 4 колонка, строки 13-14 свидетельствуют о **провалах кодификации** в ПБД ВОИС.
- Строка 6 демонстрирует пороки МПК: слишком **абстрактный уровень рубрик** международных классификаторов МПК/КПК не оставляет пользователю никакой надежды проанализировать огромное число патентов, например, рубрики B82Y30/00 «Нано-технология материалов или поверхностных эффектов, например, нано-композиты [2011.01]».
- Строка 5, колонки 4-5 позволяют предположить наличие дополнительной кодификации всемирных патентов в БПИ ЕПВ.

- Строка 6, колонки 4-5 заставляют усомниться в полноте БПИ ЕПВ в части всемирных патентов. Иначе, как получилось, что по рубрике МПК G01Q70/08 в БПИ ВОИС всемирных патентов в 10 раз больше, чем в БПИ ЕПВ даже при наличии дополнительной кодификации всемирных патентов в ЕПВ?
- Строка 9 содержит сводную информацию о **качестве кодификации патентов** в БПИ ВОИС и ЕПВ по конкретной рубрике МПК/КПК.
 - Точность классифицирования - отнесения патентов к рубрике классификатора невысока, в частности, 2 документа из 5 рубрике не соответствуют рубрике МПК C01B32/176 в БПИ ЕПВ (16 колонка), что может вообще разочаровать пользователя в классификационном поиске через индексы МПК/КПК.
 - **Полнота классифицирования** - привязки патентов к рубрике МПК аномально низка – на уровне 1%, т.е. 99 из нужных 100 пользователю патентов через классификационный поиск недоступно (10 колонка). Для экспертизы патентной заявки также важна непатентная литература, объем которой может превышать реальное число патентов в 3-5 раз (13 колонка), притом, что НПЛ в принципе через классификационный поиск недоступна. Вопрос о причинах значительной разницы в числе патентов, найденных по аналогичным сложным запросам для поиска в тексте в ПБД ЕПВ и БНТИ Гугл (колонки 11 и 13), остается нерешенным.

Выводы по материалам таблицы 1:

- Международные классификаторы МПК/КПК в виде, доступном в международных базах патентной информации ВОИС и ЕПВ, для практической работы патентного поверенного, изобретателя, конструктора, технолога, предпринимателя и инвестора непригодны. Вся экспертиза патентных заявок и действующих патентов, выполнявшаяся последние 20 лет (срок действия патента) в национальных патентных ведомствах, региональных, международных и иных патентных организациях и на предприятиях, вероятно, несостоятельна.
- Поиск в текстах в принципе может обеспечить более качественный доступ к НТИ по темам / рубрикам патентных классификаторов (высокая полнота при нормативной точности не хуже 80%). Причем, кроме патентов, доступна как непатентная литература, так и разнообразные дополнительные источники НТИ, что безусловно повышает качество доступа к НТИ, в частности, полноту материалов для анализа.
- Качественный доступ к НТИ через поиск в текстах позволяет оценить обзримость результатов поиска по конкретной рубрике официального, корпоративного или проектного классификатора с его дальнейшей детализацией (добавлением рубрик нижнего уровня). Обзримой можно считать информацию объема машинописной страницы текста (обычно не более 2000 знаков) или в худшем случае десяток документов разумного объема. Только детализованные классификации могут обеспечить приемлемое качество работы с национальными, региональными, международными и корпоративными фондами НТИ. Попытки охватить весь универсум знаний всегда проваливались за отсутствием конкретики в виде границ области знаний или решаемых проблем и портрета пользователя. Детализацию классификатора можно вести на основе исчерпывающего учета следующих источников и факторов:
 - технического задания пользователя или заказчика;
 - доступных описаний границ предметной области и опорных терминов: классификаторов, стандартов, справочников, словарей, тезаурусов, онтологий, баз знаний...;
 - коллекции глобальной НТИ на английском языке объемом более 500 млн. записей (например, БНТИ Гугл);
 - досье на пользователя (скрытые ожидания и необъявленные намерения);
 - нерешенных проблем и глобальных тенденций развития области знаний, производства или бизнеса;
 - пользовательской коллекции полезных материалов (т.н. обучение по выборке);
 - иных представленных заказчиком материалов, например, документов корпоративного документооборота.
- Качественный доступ к НТИ может обеспечить
 - исчерпывающую оценку уровня техники, в частности, избегание изобретения колеса или вечного двигателя;
 - отклонение банальных патентных заявок, выделение средств только на прорывные НИОКР,
 - обоснование господдержки и частного инвестирования оригинальных национальных производств,
 - поиск технологий для заимствования или лицензирования
 - и надежное оспаривание любых патентов, ограничивающих национальных производителей или затрудняющих продвижение отечественных товаров и услуг на зарубежные рынки.

Тестовый сложный запрос № 1 ПБД ЕПВ для поиска в текстах по теме "МПК C01B32/176 Резка углеродных нанотрубок":

((fxt="cutting CNT" OR fxt="cutting CNTs" OR fxt="cutting MWCNTs" OR fxt="cutting SWCNTs" OR fxt="cutting carbon nanotubes" OR fxt="cutting of CNT" OR fxt="cutting of CNTs" OR fxt="cutting of MWCNT" OR fxt="cutting of MWCNTs" OR fxt="cutting of SWCNT" OR fxt="cutting of SWCNTs" OR fxt="cutting of carbon nanotube" OR fxt="grinding carbon nanotubes" OR fxt="cutting of carbon nanotubes" OR fxt="disperse CNTs" OR fxt="milling CNT" OR fxt="milling CNTs" OR fxt="milling MWCNT" OR fxt="milling MWCNTs" OR fxt="milling SWCNTs" OR fxt="milling carbon nanotubes" OR fxt="milling of CNT" OR fxt="milling of CNTs" OR fxt="milling of MWCNT" OR fxt="milling of MWCNTs" OR fxt="milling of carbon nanotube" OR fxt="milling of carbon nanotubes" OR fxt="pulverizing carbon nanotubes" OR fxt="scission of CNTs" OR fxt="scission of carbon nanotubes" OR fxt="shortening CNT" OR fxt="shortening CNTs" OR fxt="shortening SWCNTs" OR fxt="shortening carbon nanotubes" OR fxt="shortening of CNT" OR fxt="shortening of CNTs" OR fxt="shortening of MWCNT" OR fxt="shortening of MWCNTs" OR fxt="shortening of SWCNT" OR fxt="shortening of SWCNTs" OR fxt="shortening of carbon nanotubes" OR fxt="slicing CNTs" OR fxt="slicing carbon nanotubes" OR fxt="slicing of CNTs" OR fxt="slicing of carbon nanotubes" OR fxt="to cut CNT" OR fxt="to cut CNTs" OR fxt="to cut MWCNTs" OR fxt="to cut SWCNTs" OR fxt="to cut carbon nanotubes" OR fxt="to shorten CNTs" OR fxt="to shorten carbon nanotubes") OR ((fxt="CNT" OR fxt="CNTs" OR fxt="MWCNT" OR fxt="MWCNTs" OR fxt="SWCNT" OR fxt="SWCNTs" OR fxt="carbon nano-tube" OR fxt="carbon nano-tubes" OR fxt="carbon nanotube" OR fxt="carbon nanotubes" OR fxt="carbon tube" OR fxt="carbon tubes") AND (fxt="chopping" OR fxt="cut of" OR fxt="cut to" OR fxt="cutting" OR fxt="mill to" OR fxt="milling" OR fxt="pulverizing" OR fxt="scission" OR fxt="sectioning" OR fxt="shorten" OR fxt="shortening" OR fxt="slicing" OR fxt="to cut") AND (fxt="average length" OR fxt="length average" OR fxt="length shortened" OR fxt="length shortening" OR fxt="mean length" OR fxt="medium length" OR fxt="short length" OR fxt="shortening length" OR fxt="uniform length"))))

Найдено **5661** патентов при поиске по всем текстовым полям, точность поиска не хуже – 80%, полноту сравнивать, увы, не с чем.

Тестовый сложный запрос № 2 ПБД ЕПВ для контроля полноты выборки запросом для поиска в тексте кодифицированных всемирных патентов для поиска в текстах по рубрике "МПК C01B32/176 Резка углеродных нанотрубок":

rp=WO src=C01B32/176 NOT (запрос № 1 см. выше)

Найдено **12** всемирных патентов из 29, кодифицированных в ПБД ВОИС по рубрике МПК C01B32/176 (5 колонка таб.1), причем все, судя по заголовку и реферату, не соответствуют указанной рубрике! Таким образом, точность кодификации в ПБД ВОИС по рубрике МПК C01B32/176 составляет (29-12)/29*100% = 59%.

Тестовый сложный запрос № 3 БНТИ Гугл для поиска в текстах по рубрике "МПК C01B32/176 Резка углеродных нанотрубок":

((("CNT" OR "CNTs" OR "MWCNT" OR "MWCNTs" OR "SWCNT" OR "SWCNTs" OR "carbon nano-tube" OR "carbon nanotubes" OR "carbon nanotube" OR "carbon nanotubes" OR "carbon tube" OR "carbon tubes") AND ("chopping" OR "cut of" OR "cut to" OR "cutting" OR "mill to" OR "milling" OR "pulverizing" OR "scission" OR "sectioning" OR "shorten" OR "shortening" OR "slicing" OR "to cut") AND ("average length" OR "length average" OR "length shortened" OR "length shortening" OR "mean length" OR "medium length" OR "short length" OR "shortening length" OR "uniform length")) OR "cutting carbon nanotubes" OR "cutting of CNT" OR "cutting of CNTs" OR "cutting of MWCNTs" OR "cutting of carbon nanotubes" OR "milling of CNT" OR "milling of CNTs" OR "milling of carbon nanotube" OR "scission of carbon nanotubes" OR "shortening of CNT" OR "shortening of CNTs" OR "shortening of MWCNTs" OR "shortening of carbon nanotubes"

Найдены **20176** патентов + **26335** НПЛ = **46511** документов, точность поиска не хуже – 80%, полноту сравнивать, увы, не с чем.

Проблемы использования искусственного интеллекта для доступа к НТИ см. в статье «Тестирование ИИ-чат-бота ChatGPT 3.5» http://onlineti.ru/download/ChatGPT.3.5_IPC.C01B32_testing.pdf. Возможности доступа к НТИ см. в обзоре «Способы получения знаний» http://www.knowledgebase.ru/Knowledge_discovery_ways.pdf.