Цитирование как средство обнаружения дополнительных знаний и их автоматического индексирования

В.В. Ежела, В.Э. Бунаков
С.Б. Луговский, В.С. Луговский, К.С. Луговский
Центр данных физики частиц “КОМПАС” в ИФВЭ, Протвино, Россия

Аннотация

В этой работе мы проверяем идею о возможности и эффективности использования формализованной пристатейной библиографии и цитируемости публикаций для автоматизированного обнаружения текстов, содержащих новые (дополнительные) знания, релевантные выделенной научной предметной области, и их автоматического индексирования метаданными избранной предметной области.

Введение

Мы исходим из того, что выделенную предметную область знаний можно отождествить с множеством научных текстов, над которым определена метаструктура, описываемая метатерминами ассоциированнымии с текстами и их связями. Предполагается также, что это множество текстов и метаструктура реализованы в вычислительной среде как набор баз данных и архивов с развитым поисковым аппаратом.

Легко видеть, что кроме терминологической структуры выделенная предметная область обладает естественной динамической ориентированной структурой - структурой гипертекста, реализуемой ссылками из текста (узла) на другие тексты (узлы) из этой же предметной области. Динамика гипертекста - это появление новых текстов, содержащих новые или дополнительные знания, релевантные выделенной предметной области, их связывание с существующими узлами и погружение экспертами в базу данных предметной области.

Релевантность нового текста выделенной предметной области традиционно определяется экспертами-индексаторами после ознакомления с текстом. Индексаторы определяют потенциальное место текста в структуре знаний предметной области, фиксируя это место приписыванием тексту метатерминов. Окончательно место текста определяется после фактографической обработки и уточнения набора приписанных индексаторами метаданных. Это трудоемкий и дорогостоящий процесс.

Между тем место нового текста в структуре гипертекста уже в какой-то степени определено авторами нового текста, которые дают ссылки на использованные знания из других текстов. Такие ссылки в грамотно подготовленных текстах выделены в специальный раздел и записываются по международным стандартам.

Таким образом, возникает возможность автоматического определения потенциального места нового текста в структуре знаний предметной области путем приписывания новому тексту метаданных цитируемых им текстов.

Реальное место прединдексированного текста в базе данных (реально приписанные ему метаданные) определится экспертом, речь идет лишь о создании механизма предварительного отбора и индексирования текстов.

Идея автоматизированного индексирования на основе сетей цитирования многократно обсуждалась и даже реализована в некоторых программных продуктах [1], [2], [3], [7], [8].

Вклад настоящей работы состоит в практической реализации и тестировании метода применительно к конкретной предметной области - физике частиц, которая благодаря своим особенностям (наличие оперативно пополняемых библиографических и фактографических баз данных, полнотекстовых архивов) предоставляет хороший полигон для библиометрических исследований.

На основе фактографических баз данных физики частиц, сопровождаемых в центре данных ИФВЭ и Лаборатории Лоуренса [4], [5], и библиографической базы данных HEP(USPIRES) [6], сопровождаемой библиотеками SLAC и DESY, построена модель процедуры автоматического индексирования по цитированию. Произведены системные измерения цитирования различных тематических разделов Review of Particle Physics [4] из текстов, каталогизированных в базе данных HEP, получены оценки эффективности предложенного метода, обсуждаются его достоинства и недостатки.

Цитометр: отбор публикаций релевантных RPP

Для компактности изложения нам понадобятся обозначения для разных частей текста. Будем отождествлять текст издателя τ и его образ T в какой-либо компьютерной информационной системе (КИС). Нам достаточно будет рассматривать текст как многокомпонентный объект:

$$T = (R, A, B, L, M^1, M^2, ..., M^k),$$

где:

R - набор библиографических атрибутов в формате издателя и соответствующие компьютерные коды в КИС, URL;

A - оригинальная аннотация;

B - “тело” текста от введения до заключения;

L - формализованный (т.е. готовый к компьютерному использованию) список цитированной литературы;

$M^1, ..., M^k$ - метаданные в системе издателя (PACS, например) и наборы метаданных в КИС.

В нашем анализе мы будем использовать следующие КИС:

- База данных HEP: $(R, A, L, M^{SLAC}, M^{DESY})$
- Сайт издателя prola.aps.org (PROLA): (R, A, B, L, M^{PACS})
- База данных ISI: (R, L, M^{ISI})

- База данных RPP: $(R, M^{RPP} : DATA, RVUE)$

Будем оценивать эффективность использования инфоресурсов HEP, PROLA, ISI и метода “автоматического” отбора и индексирования, основанного на цитировании (метод Цитометра), сравнивая полученные результаты с результатами отбора и индексирования проведенного экспертами из PDG(LBL).

Характеристики библиографической части RPP в сравнении с HEP (на июль 2001):

Таблица 1.	Статистика отражения публикаций в HEP и RPP
11818	$\lvert RPP \rvert$ - число записей о публикациях по физике частиц в RPP
10340	$\lvert RPP_R \rvert$ - число записей в RPP о публикациях в реферируемых журналах
8929	$\lvert RPP \bigcap HEP \rvert$ - число записей в пересечении HEP и RPP
6325	$(T \in RPP \bigcap HEP, L \neq \varnothing)$

Как указано во введении, мы исходим из того, что релевантность текста предметной области базы знаний RPP в подавляющем большинстве случаев определяется фактом цитирования опубликованного обзора RPP, сайта RPP, или цитирования текстов T', уже использованных ранее в обзоре RPP, $(T' \in RPP)$.

Для проверки этой гипотезы и оценки доли ее “отказов” мы приводим распределение числа текстов из RPP по числу цитирований каждого из них другими текстами из RPP. Из 6325 текстов в RPP, для которых в HEP имеются данные о цитированной ими литературе, только 224 не имеют социтирований с RPP.

Таким образом, можно оценить долю потерь новых (дополнительных) знаний по методике цитометра на HEP как 3.5%. Интересно отметить, что положение максимума в распределении приходится на число цитирований “из RPP в RPP”, равное 4-5. Это отражает сложившуюся в экспериментальной физике частиц практику: явление приобретает статус открытия (надежно установленного знания) после не менее пятикратного его подтверждения и уточнения.

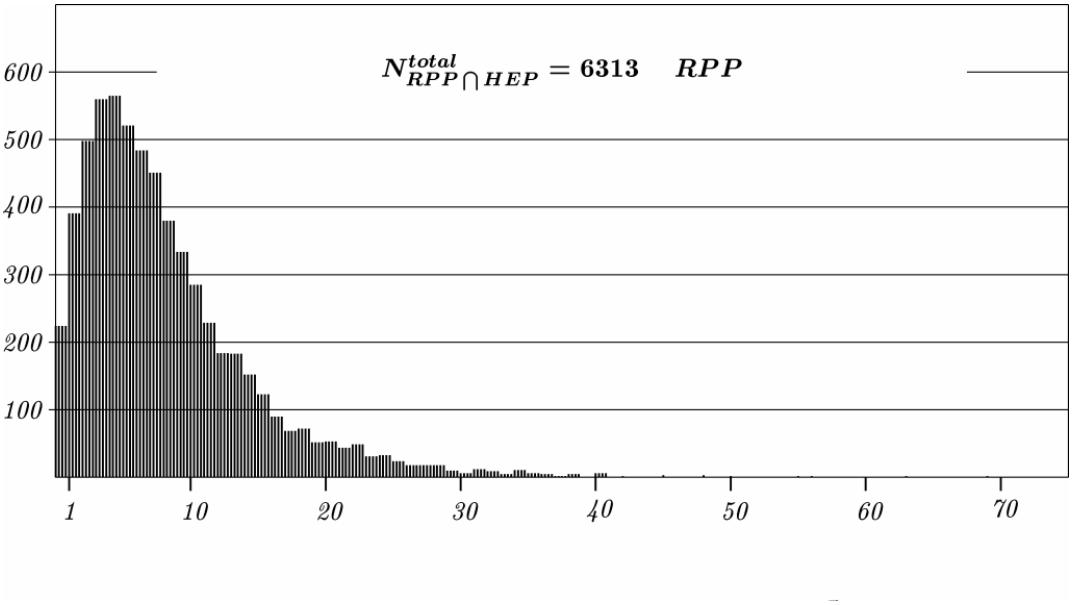

Рис.1: Распределение $N^{total}_{RPP \bigcap HEP} = 6313$ по цитируемости из RPP

Для оценки эффективности метода цитометра со стороны трудозатрат по отбраковке текстов, содержащих цитирования обзоров RPP, и публикаций из базы данных RPP (социтирования RPP), но не содержащих сведений, полезных для отражения в базе RPP или в миниобзорах, посмотрим на динамику цитирований изданий RPP в HEP и на динамику экспертного отбора публикаций для RPP.

Опишем, сначала традиционный отбор. База данных RPP сопровождается усилиями экспертов-физиков из нескольких организаций. Отбор литературы производится путем "глазного" сканирования оглавлений журналов с принятием решений "брать – не брать" на основе знаний эксперта. Если данных из названия не хватает, то просматривается и авторская аннотация, если и ее не хватает, то эксперт просматривает весь текст статьи.

Разметка тематики статьи производится приписыванием RPP-метаданных к библиографической записи. Метаданные в RPP - это коды имен физических объектов и имен наблюдаемых (измеряемых) величин, всего более 5000 кодов. Просмотрщик литературы приписывает к библиографии статьи коды имен частиц, о которых могут быть сведения в статье, и/или коды релевантных миниобзоров из RPP, всего около 600 терминов.

Эту технологию мы называем экспертным отбором, или экспертным индексированием. В таблице 2. приведены статистики заполнения баз данных HEP и RPP путем экспертных отборов и "пробы" простейшего варианта отбора релевантных базе RPP публикаций по цитированию опубликованных обзоров RPP в публикациях, учтенных в базе HEP.

Таблица 2.	Записи в HEP о статьях в журналах			
$Year(Yr)$	$N^{HEP}_{\le Yr}$	$N^{RPP}_{\le Yr}$	$C^{RPP}_{\le Yr}$	$\in RPP \le Yr$
1974	3545	2584	95	10
1975	8314	2827	234	37
1976	13878	2998	346	59
1977	19675	3205	429	74
1978	25545	3442	507	85
1979	31847	3623	610	96
1980	37864	3847	719	113
1981	44358	4092	854	134
1982	50731	4320	969	160
1983	56719	4565	1071	181
1984	63771	4813	1197	213
1985	70459	5046	1287	236
1986	76506	5313	1410	259
1987	84077	5596	1672	318
1988	90927	5887	1883	377
1989	99506	6160	1989	427
1990	107832	6565	2521	526
1991	116611	6917	2928	620
1992	126437	7353	3382	746
1993	134788	7723	3703	877
1994	145222	8024	4240	977
1995	154996	8458	5125	1144
1996	165021	8968	5911	1332

1997	175891	9468	6852	1543
1998	189377	10062	7789	1786
1999	205139	10586	8783	1977
2000	218104	11158	9500	2198

Сравнение колонок 2 и 3 таблицы 2 показывает, что экспертный отбор к 2001 году стал довольно жестим фильтром. Отбираются только 5-6% от общего числа публикаций по физике частиц и смежным дисциплинам, которые и образуют массив для пополнения фактографической базы данных по свойствам частиц и тематических обзоров их взаимодействий. Сравнение колонок 4 и 5, где приведены данные кумулятивного цитирования бумажных выпусков RPP из журнальных публикаций, показывает, что фильтр, основанный на цитировании только опубликованных обзоров RPP, более жесткий, чем экспертный фильтр и он сильно "перекошен", пропускает только 4.4% публикаций, среди которых только 23% пертинентны RPP.

Таким образом, мы получаем подтверждение того, что цитирований только обзоров RPP явно недостаточно. Нужно учитывать и цитирования текстов отраженных в RPP. Требование цитирования публикаций из базы RPP необходимы не только для проведения автоматического прединдексирования, но также и для обеспечения полноты отбора релевантных публикаций.

Для отбора релевантных RPP публикаций на основе социтирований с RPP был создан специализированный полуавтоматический фильтр – "цитометр," который мы возможно опишем в другой работе, а здесь приведем результаты его обкатки на реальных данных.

С помощью цитометра были отобраны все записи базы данных HEP о публикациях в журналах за 2000 год. Результаты классификации и отбора для экспертного просмотра сведены в таблицу 3.

Таблица 3.		HEP (Year = 2000, published) :		**12792**
$L = \varnothing$	**3498**	L≠∅		**9294**
		$L \bigcap RPP \neq \varnothing$: **3400**	$L \bigcap RPP = \varnothing$:	**5894**
		Mini: **1249**	Exp : **2151**	
$R \in RPP$	**83**	**24**	**361**	**4**

Таким образом, цитометр предлагает для экспертного просмотра 3498+3400=6898 публикаций вместо 12792, с потерей 4 из 472 публикаций, найденных экспертами при "тотальном" просмотре в 2000 году. При этом к 3400 публикациям из 6898 будут автоматически приписаны RPP-метаданные (по социтируемости, см. следующий раздел). Из 472 публикаций, отобранных экспертами, 385 будут иметь автоматически приписанные RPP -метаданные.

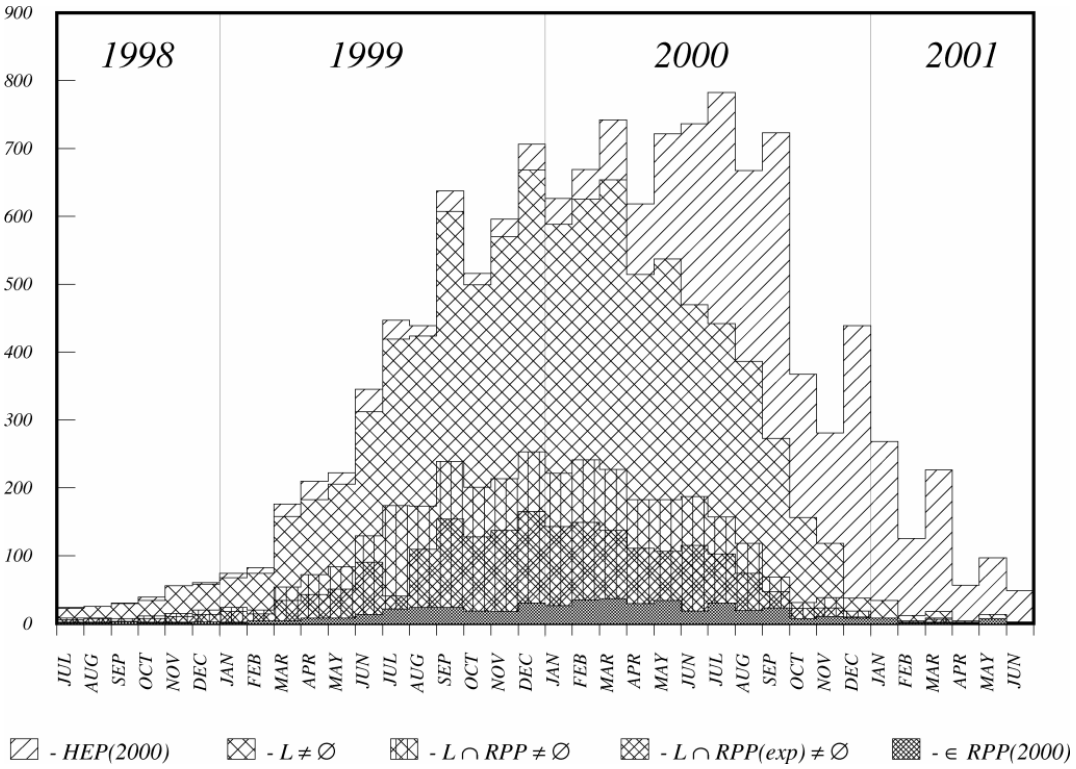

Рис. 2: Распределение выборки и её “фильтраций” по интервалам пополнения базы HEP.

Из-за затрудненности доступа к базе данных HEP (медленная связь, ограничения объема одноразовой выборки, ограничения по использованию WEB-роботов) мы использовали доступ через электронную почту с ее разборкой “домашним роботом”. Для сокращения объемов пересылок пришлось дробить выборку по ежедневным дополнения в базу HEP. Распределения полученной “полной” выборки и результатов ее “фильтрации по социтируемости” по месяцам дополнения представлены на рисунке 2. Распределения могут быть полезными в планировании работы по экспертной обработке отфильтрованной выборки.

Полученные данные уже можно использовать для оценок эффективностей отбора релевантных публикаций цитометром и экспертами.

- Эффективность экспертного отбора: $\frac{100 \cdot (12792 - 472)}{12792} = 96\%$

- Эффективность отбора цитометром можно оценить, если предположить, что при дополнении записей из HEP недостающими списками цитированной литературы (класс с $L = \varnothing$) доля поправленных записей с $L \bigcap RPP \neq \varnothing$ будет такой же как и в классе с $L \neq \varnothing$. Тогда эффективность цитометра получается 61%

Как следует из таблицы 3 и рисунка 2, если бы отбирались толко публикации дающие цифровые данные, то эффективность их автоматического отбора была бы выше -- 73%. По-видимому, это связано с более точным цитированием работ содержащих экспериментальные данные.

Таким образом, фильтр, основанный только на социтировании публикаций и базы данных RPP оказывается недостаточно эффективным в отборе релевантных публикаций. И здесь, по-видимому, есть резерв. Необходимо проверить, как сработает фильтрация по цитируемости отобранных по социтированию с RPP текстов. Это мы планируем сделать в скором будущем. Однако фильтр по социтируемости может быть только "рекомендательным", но не отсекающим, так как цитометр предполагается использовать для индексирования новых текстов сразу после их появления. По-видимому, для повышения эффективности отбора не обойтись без использования метаданных. К сожалению, в базе данных HEP, где есть информация о цитировании и может быть вычислена информация о цитируемости, аппарат предметного индексирования беден, и база проиндексирована фрагментарно. Кроме того, техническое ограничение на длину запроса не позволяет сформулировать отсекающие ограничения в одном запросе без больших потерь. Физика частиц сильно перекрывается, например, с ядерной физикой, физикой космических лучей, астрофизикой и др., и в базу данных HEP (High Energy Physics) с неизбежностью попадает большая часть текстов из этих смежных областей, не релевантных физике частиц, а отсекать их по грубому тематическому рубрикатору, используемому в HEP, не удается. Возникают большие потери релевантных текстов. Использование рубрикатора PACS могло бы помочь делу, однако записи в HEP не проиндексированы PACS-кодами.

Итак, проблема автоматизации индексирования метаданными узкой предметной области сильно переплетается с проблемой автоматизации индексирования по рубрикаторам.

Автоматическое индексирование

Имея результаты отбора релевантных публикаций и автоматически приписанные им RPP-метаданные, мы можем оценить "точность" и эффективность автоматического индексирования.

RPP-метаданные приписывались к библиографической записи по следующему правилу: если социтирование публикации T и базы данных RPP не пусто, то $M^{RPP}(T) = \bigcup_{T'} M^{RPP}(T')$ для всех $T' \in RPP$ таких, что $L(T) \bigcap R(T') \neq \varnothing$. При этом подсчитывались кратности приписываемых метаданных. Сравнение наборов метатерминов (с их кратностями), приписанных цитометром к 385 текстам, ранее отобранным и проиндексированным экспертами из PDG, показывает, что в 325 текстах наборы метаданных цитометра и наборы экспертных метаданных имеют непустые пересечения. В 197 случаях экспертные метатермины совпадают с метатерминами цитометра с наибольшей кратностью.

В 60 (из 385) текстах наборы метаданных цитометра и экспертов не пересекаются. Беглое рассмотрение всех этих случаев показало, что 9 из них связаны с еще не устраненной синонимией и избыточностью в метаданных RPP, описывающих массивные нейтрино и нейтринные осцилляции, 17 - с появлением нового метатермина "extra dimentions" и связанной этим синонимией в метаданных, внесенной нами при формировании таблиц соответствия. Остальные 34 еще требуют более углубленного анализа полного текста статей. Вероятны и механические описки экспертов, внесенные при индексировании. (Напомним, что мы анализируем только тексты, отобранные как релевантные по анализу названий статей и еще не прошедших фактографическую обработку).

Таким образом, эффективность автоматического индексирования можно оценить как $\frac{325 \cdot 100}{385} = 84.4\%$, а точность индексирования метатермином наибольшей кратности оценивается в 90% при условии, что наибольшая кратность >3.

Обратимся еще раз к рисунку 2. Соотношение правых склонов распределений показывает, что большая часть записей в HEP с $L = \varnothing$ как раз лежит в области правых склонов, т.е. это записи дополненные во второй половине 2000 и в начале 2001 гг.

Явно имеется запаздывание с дополнением записей информацией о цитируемой в соответствующих текстах литературе. По-видимому, это естественное запаздывание есть в любой КИС, содержащей цитирование. Оно обусловлено необходимостью поправок многочисленных описок авторов и нормализации форматов записей литературных ссылок к используемому в КИС. В HEP это, так называемый, стандарт SLAC-CODEN.

Если нужна оперативность, то естественно обратиться к сайтам издателей и работать непосредственно с информационным ресурсом издателя. Однако издатели, как правило не дают легкого доступа к информации о цитированной в статье литературе. Приятным для физиков исключением являются сайты APS: publish.aps.org и prola.aps.org, где выкладываются все данные:

$$T = (R, A, B, L, M^{PACS})$$

причем, библиографические атрибуты, абстракт, метаданные и использованная литература доступны для "перекачки" в форматах удобных для обработки программами.

Естественно наше желание использовать этот важный для физиков инфоресурс для обкатки методики цитометра, сначала на метаданных издателя -- PACS-кодах и затем на проблемно ориентированных метаданных. В случае успеха мы надеемся получить возможность автоиндексирования используемых информационных ресурсов кодами единого тематического рубрикатора, к необходимости которого мы пришли в конце второго раздела. Мы будем обсуждать в этой работе только метаданные издателя. Т.е. с помощью цитометра попытаемся проставить PACS-коды (будем обозначать их CPACS) для публикаций и сравнивать их с PACS-кодами, проставленными авторами под контролем редакции и издателя. Результаты сравнения позволят нам сделать оценки как инфоресурса, так и эффективности цитометра на нем. В связи с известными ограничениями технического и организационного характера нам удалось обработать, только малую толику публикаций и потому результаты весьма предварительные и могут дать смещенные оценки. Но и они получаются достаточно интересными.

Из 4067 записей типа $T = \left(R, L, M^{PACS}\right)$, отобранных роботом с сайта издателя, только 1956 оказались пригодными для применения методики цитометра. Основная причина это ограниченность выборки.

К каждой записи из этой выборки $T = \left(R, L, M^{PACS}\right)$ цитометром были приписаны CPACS-коды из цитируемых и представленных в исходной выборке (1956) статей с их кратностями:

$$\left(R, L, M^{PACS}\right) \rightarrow Цитометр \rightarrow \left(R, L, M^{PACS}, M^{CPACS}\right)$$

и анализировалось соотношение максимальных кратностей приписанных цитометром PACS-кодов:

- m^+ в пересечения $M^{PACS} \bigcap M^{CPACS}$
- m^- в разности $M^{CPACS} \setminus M^{PACS}$

в каждой записи. По основной идее цитометра если $m^+ > m^-$, то цитометр успешно восстанавливает метаданные экспертов (авторы, рецензенты, редакторы).

Статистика такова:

$$m^+ > m^- \to 813$$
$$m^+ = m^- \to 716$$
$$m^+ < m^- \to 427$$

Таким образом, в 813 случаях цитометр индексирует так же, как эксперты, в 427 случаях цитометр индексирует "неверно". В 716 случаях необходима дополнительная информация. На первый взгляд ситуация для цитометра (или для PACS-рубрикатора) весьма плачевна. Однако давайте посмотрим на структуру PACS-кодов, по фрагменту PACS-рубрикатора:

.....

14. Properties of specific particles

14.20.-c Baryons (including antiparticles)
14.20.Dh Protons and neutrons
14.20.Gk Baryon resonances with S=0
14.20.Jn Hyperons
14.20.Lq Charmed baryons
14.20.Mr Bottom baryons
14.20.Pt Dibaryons

14.40.-n Mesons
14.40.AQ Pi, K, and eta mesons
14.40.Cs Other mesons with S=C=0, mass < 2.5 GeV
14.40.Ev Other strange mesons

....

Как видно из фрагмента, PACS это трехуровневый рубрикатор. Редакция издателя рекомендует использовать не более 4-х кодов, и по возможности самого нижнего уровня. Для информативных политематических отчетов эти ограничения приводят к вынужденным приписываниям не всех рубрик низшего уровня, необходимых для тематического "покрытия" результатов в тексте, и к несоответствиям в отобранных узких рубриках статьи и рубриках цитируемых статей. Эти ограничения ослаблены для обзорных статей в Reviews of Modern Physics.

Если проводить сравнения PACS и CPACS по старшим двум уровням, то статистика соответствия индексирований цитометра и экспертов улучшается:

$$m^+ > m^- \rightarrow 1078$$
$$m^+ = m^- \rightarrow 566$$
$$m^+ < m^- \rightarrow 312$$

Для самого верхнего уровня она такова:

$$m^+ > m^- \rightarrow 1508$$
$$m^+ = m^- \rightarrow 253$$
$$m^+ < m^- \rightarrow 195$$

Улучшение точности автоиндексирования с 41% до 77% дает основание полагать, что с увеличением статистики в игру вступят цитирования и PACS-ы других APS-журналов: Phys.Rev.Lett, Phys.Rev.A,B,C,E, Rev.Mod.Phys. и метод цитометра заработает более успешно и на нижнем уровне. Однако напомним, что в проведенном сравнении использовались не все ссылки из L, а только на отобранные выпуски Phys.Rev.D58,-D63. Это связано с отсутствием элемента M^{PACS} у многих статей из европейских и российских журналов, а доля ссылок на них в каждом L из отобранных с сайта prola.aps.org более половины. Поэтому полученные весьма предварительные данные дают смещенные оценки, которые пока нельзя использовать для надежных выводов.

Заключение

Полученные результаты использования идеи цитометра в автоматизации обнаружения опубликованных текстов релевантных выделенной предметной области и их автоматическое индексирование метаданными предметной области подтверждают состоятельность идеи для системы “разделяющих” метаданных. (на примере метаданных системы RPP). Результаты автоматического индексирования наиболее точны (совпадают с результатами экспертного индексирования) если используется вся информация о социтировании статей в избранной предметной области. Это накладывает дополнительные требования к полноте и однородности в заполнении информацией используемых в цитометре информационных полей библиографических КИС.

Как мы видели в HEP в более чем 30% записей о журнальных публикациях информация о цитируемых ими работах не содержится. Это определенно вносит искажения в результаты цитометра, поэтому для формирования окончательных выводов о состоятельности цитометра необходима контрольная проверка эффективности и точности цитометра на базе ISI и метаданных RPP. Мы планируем сделать это, как только получим достаточно технологичный доступ к базе ISI.

Полученные данные о “результативности” цитометра на данных сайта PROLA и PACS-кодах хотя и не опровергают идею еще очень скудны для содержательных выводов. Для корректного использования необходимо использовать весь массив PROLA после 1975 года (года внедрения в APS классификации PACS), и проиндексировать все цитируемые публикации других издателей PACS-кодами. Это приводит к необходимости дальнейших системных измерений, и подключения не только цитирования, но и цитируемости текстов, для повышения точности автоиндексирования. При достаточно высокой точности автоиндексирования можно будет пытаться решать задачи интегрирования научных инфоресурсов и онтологий на более или менее твердой базе – сетях цитирования.

Благодарности

Работа выполнена при частичной поддержке фонда РФФИ, гранты 01-07-90392, 01-07-90432. Авторы благодарят сотрудничество PDG, библиотеки SLAC и DESY за предоставленную возможность использования в этой работе их информационных сетевых ресурсов. Мы благодарны также администраторам сайта “PROLA,” за предоставленную возможность использования сетевого робота для ``перекачки" необходимых данных.

Список литературы

[1] R.D. Cameron. A Universal Citation Database As a Catalyst for Reform in Scholarly Communication. First Monday, April 1997.
http://www.firstmonday.dk/issues/issue2_4/cameron/index.html

[2] S. Lawrence, K. Bollacker, C. Lee Giles. Digital Libraries and Autonomous Citation Indexing. IEEE Computer, Vol.32, No.6, pp.67-71, 1999.

[3] K. Bollacker, S. Lawrence, C. Lee Giles. A System for Automatic Personalized Tracking of Scientific Literature on the Web. Proceedings of the Fourth ACM Conference on Digital Libraries, ACM Press, New York, pp.105-113, 1999

[4] G. Groom et al. Review of Particle Physics. The European Physics Journal C15, 1
http://dbserv.ihep.su/pdg/index.html

[5] COMPBS Group. Particle Physics Data System.
http://wwwppds.ihep.su:8001/ppds.html

[6] SLAC & DESY libraries. HEP(USPIRES) database.
http://www-spires.slac.stanford.edu/find/hep

[7] Science Citation Index.
http://www.isinet.com/isi/products/citation/sci/index.html

[8] Eugene Garfield homepage.
http://www.garfield.library.upenn.edu/

[9] Physics and Astronomy Classification Scheme.
http://publish.aps.org/PACS/

[10] Издательские сайты APS.
http://publish.aps.org/
http://prola.aps.org/

DISCOVERY OF THE ADDITIONAL KNOWLEDGE AND THEIR AUTOMATIC INDEXING VIA CITATIONS

V.V.Ezhela, V.E.Bunakov, S.B.Lugovsky, V.S.Lugovsky, K.S.Lugovsky
Particle Physics Data Center “COMPAS” at IHEP, Protvino, Russia

In this work we test the idea of the possibility and effectiveness to use the lists of formalized (standardized) cited literature to automatic discovery of the texts containing the new (or additional) knowledge, relevant the given scientific subject and moreover to automatic indexing by metadata of the corresponding field of scientific knowledge. It turned out that such a possibility and effectiveness strongly depend on the quality of the information resources used.