УДК 004.021, 004.047, 004.622

ПОСТРОЕНИЕ НАУЧНЫХ ПРОФИЛЕЙ УЧАСТНИКОВ НАУЧНО - ОБРАЗОВАТЕЛЬНОГО ПРОЦЕССА В ИНФОРМАЦИОННОЙ СИСТЕМЕ УНИВЕРСИТЕТА

Варенников Д.А.[1], Шлей М.Д.[1], Муромцев Д.И.[1]

[1]Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Санкт-Петербург, Российская Федерация (197101 Российская Федерация, г. Санкт-Петербург, Кронверкский проспект, 49), e-mail: varenikovda@gmail.ru

В статье описаны подходы к построению научных профилей пользователей в информационной системе университета. Представлена информационная модель научной деятельности обучающихся, которая описывает процессы взаимодействия участников и результаты их научной деятельности. Уделено внимание вопросу идентификации авторов публикаций, участников научно-образовательного процесса, а также интеграции с авторскими профилями из наукометрических баз данных. Методы идентификации авторов публикаций позволяют повысить качество наполнения базы знаний университета по сведениям из различных наукометрических баз данных. Предложенные методы лежат в основе разрабатываемого рекомендательного сервиса для пользователей информационной системы университета. Данный сервис позволяет формировать рекомендации по выбору научного руководителя или обучающегося, научного мероприятия, грантов, публикаций и периодических изданий для публикации научных результатов.

Ключевые слова: научные интересы, научная деятельность, обучающиеся, наукометрические базы данных.

BUILDING SCIENTIFIC PROFILES FOR SCIENTIFIC AND EDUCATIONAL PROCESS PARTICIPANTS IN IT SYSTEM OF THE UNIVERSITY

Varenikov D.A.[1], Shley M.D.[1], Muromtsev D.I.[1]

[1]ITMO University, Sankt-Peterburg, Russian Federation (197101 Russian Federation, Sankt-Peterburg, Kronverkskiy prospekt, 49), e-mail: varenikovda@gmail.ru

In this paper authors describe approaches to build scientific user profiles in IT system of the University. The information model of student scientific activity is presented. This model describes the participant interaction process and their scientific activity results. Attention is paid to the identification of the publication authors when they are scientific and education process participants, to the integration with authors' profiles of scientometric databases. Author identification methods allow to improve the quantity of filling the University knowledge base using information from different scientometric databases. Proposed methods are the basis of the recommendation service for IT system users. This service gives recommendations for choosing a research adviser, students, scientific events, academic grants, publications and periodicals.

Keywords: research interests, research activities, students, scientometric database.

В настоящее время проводится активная модернизация системы высшего образования с целью повышения ее качества, в частности, значительное внимание уделяется увеличению доли научной составляющей в образовательном процессе, вопросам интеграции научной, инновационной и образовательной деятельностей, а также развитию научной деятельности в целом**.** Данный процесс сопровождается резким ростом требований ко всем участникам научно-образовательного процесса – преподавателям, магистрантам и аспирантам [11]. Среди основных требований можно выделить увеличение объема научных исследований в работе, число публикаций, активное участие в научных мероприятиях, количество полученных патентов и выигранных грантов [5]. Изменения, проводимые в деятельности вуза, требуют внедрения современных автоматизированных систем управления образовательными процессами и создания инфраструктуры для поддержки деятельности сотрудников и

обучающихся. Эффективность реализации процессов научной деятельности определяется качеством решения следующих задач: поиск и обеспечение доступа к научно-исследовательским работам, выполняемым по схожим тематикам, своевременное информирование о проведении научных мероприятий и планирование участия в них, обеспечение возможности публикации полученных научных результатов в высокорейтинговых изданиях.

Постановка задачи

Для анализа процесса реализации научной деятельности обучающихся (магистранты и аспиранты) и преподавателей вуза была построена информационная модель верхнего уровня, представленная на рисунке 1. Данная модель описывает процессы взаимодействия участников и результаты их научной деятельности в информационной системе университета.

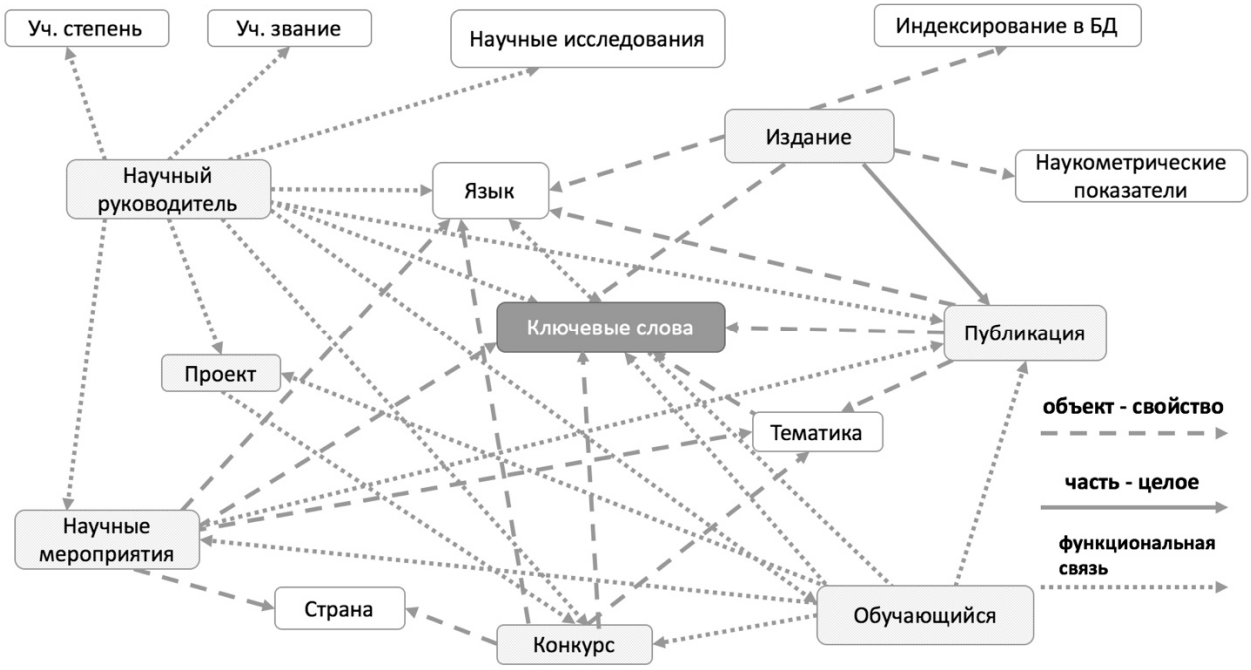

Рис. 1. Информационная модель научной деятельности

Анализ модели и предметной области позволил выделить основные проблемные области, связанные с принятием решений и поиском соответствующей информации:

- Выбор научного руководителя или обучающихся с учетом научных интересов.
- Поиск подходящих периодических изданий для публикации научных результатов.
- Выбор актуальных научных мероприятий для участия.
- Выбор конкурсов и грантов для финансирования проводимых научных исследований.

В университете организована структура проектных менеджеров, которые регулярно отслеживают новости об открываемых конкурсах и фондах и отбирают те из них, которые могут заинтересовать преподавателей и обучающихся [12].

- Поиск значимых публикаций по заданным тематикам.

В построенной модели можно выделить следующие ключевые объекты:

- Научные руководители и обучающиеся – пользователи информационной системы.
- Научные мероприятия – конференции, конгрессы, семинары, круглые столы и прочие.
- Конкурсы – конкурсы, ориентированные на материальную поддержку преподавателей и обучающихся.
- Публикации – статьи, опубликованные в периодических издания.
- Издания – периодические издания.

В представленной модели, основные информационные объекты, связаны через ключевые слова, которые формируют профиль объекта или так называемую «область научных интересов». Использование методов частотного анализа и информационных технологий позволяют выполнять анализ связей между научными интересам участников, проводимыми исследованиями, актуальными конкурсами и мероприятиями. По результатам анализа формируются рекомендации для решения вопросов, озвученных выше.

Особенно важной задачей при использовании данного подхода является формирование научного профиля пользователя информационной системы и таких объектов как издания, публикации, конкурсы, мероприятия. Поскольку от полноты информации о профиле будет зависеть качество и точность формируемых рекомендаций, а следовательно эффективность развития научной деятельности вуза. Решение данной задачи имеет свои особенности, так как при формировании профиля используется множество источников информации, и при анализе информации необходимо правильно определять приоритеты тех или иных научных интересов.

В данной статье рассмотрен процесс формирования научного профиля для участника научно-исследовательской деятельности вуза, а также других связанных информационных объектов на примере научного профиля издания.

Формирование научного профиля пользователя

Формирование научных интересов пользователя в информационной системе университета происходит за счет самостоятельного ввода информации при заполнении личного профиля в информационной системе, а также автоматического сбора сведений о ключевых словах. Автоматический сбор сведений основывается на формализации и последующей интеграции информации из наукометрических баз данных, анализа поведения пользователей в информационной системе, его научно-практических результатов на основе методов частотного анализа и алгоритмов нечеткого поиска (см. рис. 2). Под наукометрическими базами данных понимают библиографические и реферативные базы данных, а также инструмент для отслеживания цитируемости научных статей [6].

Рис. 2. Информационная модель научных интересов пользователя в информационной системе университета (ИСУ)

Автоматическое наполнение профилей ключевыми словами позволяет значительно расширить выборку, на которой в дальнейшем будет основываться инструмент предоставления рекомендаций по поставленным задачам, и повысить качество полученных рекомендаций.

Научные интересы пользователя – это множество его ключевых слов K.

$$\mathrm{K} = \mathrm{K^p} \cup \mathrm{K^a},$$

где K^p – множество ключевых слов, указанных пользователем, а K^a – множество ключевых слов, автоматически выбранных с учетом частоты их появления.

$$\mathrm{K^a} = \{\mathrm{q} \in \mathrm{Q} : \mathrm{f(q)} > c\},$$

где Q – множество автоматически полученных ключевых слов, $f(q)$ – частота появления ключевого слова q, c – пороговое значение для частоты появления ключевого слова.

$$\mathrm{Q} = \bigcup_{\mathrm{i}=1}^{h} \mathrm{Q_i},$$

где h – количество источников, на основании которых формируется множество Q. Множество Q формируется за счет:

- Посещения пользователем информационной системы. Q_1 – множество ключевых слов, полученных по результатам посещения страниц, содержащих ключевые слова и тематики;
- Анализа схожих интересов между пользователями, посетившими одинаковые страниц. Q_2 – множество ключевых слов пользователей со схожими интересами [3, 4];
- Анализа тематик публикаций, автором которых является пользователь. Q_3 – множество ключевых слов, полученных на основе анализа публикационной активности пользователя;
- Анализа схожести интересов между соавторами публикаций, автором которых является пользователь. Q_4 – множество ключевых слов соавторов публикаций пользователя;

- Получения сведений о подписки пользователя на рассылку в информационной системе. Q_5 – множество ключевых слов, указанных пользователям для получения рассылки в информационной системе;
- Анализа профиля пользователя в наукометрических базах данных. Q_6 – множество ключевых слов пользователя, полученных из наукометрических баз данных.

$$Q_1 = \bigcup_{j=1}^{m} L_j,$$

где L_j – множество ключевых слов $j-$ой страницы, m – количество страниц. В статье рассматриваются страницы информационной системы, которые посещает пользователь.

$$Q_2 = \left\{ K_v : v \in \bigcup_{j=1}^{m} V_j, Jaccard(K_v, K^p) > s \right\},$$

где K_v – множество ключевых слов пользователя v, V_j – множество пользователей, посетивших $j-$ую – страницу, за исключением рассматриваемого пользователя, $\text{Jaccard}(K_v, K^p) = \frac{|K_v \cap K^p|}{|K_v \cup K^p|}$ – мера Жаккара ($0 \leq Jaccard(K_v, K^p) \leq 1$), а s – пороговое значение схожести.

$$Q_3 = \bigcup_{j=1}^{g} Kp_j,$$

где Kp_j – множество ключевых слов $j-$ой публикации, g – количество публикаций пользователя.

$$Q_4 = \left\{ K_{ca} : c_a \in \bigcup_{j=1}^{g} Ca_j, Jaccard(K_{ca}, K^p) > s \right\},$$

где K_{ca} – множество ключевых слов пользователя p_{ca}, Ca_j – множество соавторов публикации j, за исключением рассматриваемого пользователя, $Jaccard(K_{ca}, K^p)$ – мера Жаккара, а s – пороговое значение схожести.

$$Q_6 = \left\{ K_{apr} : a_{pr} \in \bigcup_{j=1}^{d} Apr_j \right\},$$

где K_{apr} – множество ключевых слов пользовательского профиля a_{pr}, Apr_j – множество авторских профилей наукометрической базы данных j для рассматриваемого пользователя, d – количество наукометрических баз данных. В статье рассматриваются наиболее распространенные наукометрические базы данных и их идентификаторы авторских профилей:

1. РИНЦ (российский индекс научного цитирования) – используется уникальный идентификатор SPIN-код [9];

2. Web of Science – самая авторитетная в мире база данных по научному цитированию института научной информации (Institute of Scientific Information - ISI) – используемый уникальный идентификатор ResearcherID [9];

3. Scopus – это крупнейшая в мире единая мульти дисциплинарная реферативная база данных, представляющая уникальную систему оценки частоты цитирования. Используемый уникальный идентификатор ORCID [9];

Авторский профиль из наукометрических базах данных в информационной системе представлен следующим образом:

$$Apr = \langle K_{apr}, P_{pr}, Ind_{pr} \rangle,$$

где P_{pr} – множество публикаций авторского профиля,Ind_{pr}– идентификатор авторского профиля.

Анализ ключевых слов K_{apr} авторских профилей Apr, полученных из наукометрических баз данных, начинается с определения связей между пользователями информационной системы (авторами публикаций) A и наукометрическими базами данных. Множество авторов публикаций в информационной системе представлено следующий образом:

$$A = \{a_i\}_{i=1}^{z},$$

где z – количество уникальных авторов.

Определение связей авторских профилей, полученных с наукометрических баз данных, и пользователями информационной системы является первостепенной задачей. Один из возможных подходов идентификации авторов публикаций из различных баз данных публикаций – это проведение анализа возможных внешних идентификаторов авторов и сопоставление их с внутренними (университетскими) идентификаторами [1]. Такие связи идентификаторов не всегда существуют, возникают новые авторские коллективы, автор может изменить фамилию, также в авторитетных базах данных авторы могут не иметь уникальный идентификатор, или один и тот же автор может быть связан с разными идентификаторами. В настоящее время в мире нет единого стандартизованного способа идентификации журнальных статей, авторов, их мест работы и др., несмотря на то, что в последние годы введены в действие немалое число различных идентификаторов [8]. При идентификации авторов большое значение имеет аффилиация. Некоторые авторы не указывают аффилиацию с университетом, что приводит к затруднению их идентификации. В случае работы с аффилиациями можно выделить следующие возможные варианты:

- Указана аффилиация – автор является сотрудником университета и указал ссылку на университет [1].
- Отсутствие аффилиации – автор является сотрудником университета и не указал ссылку на университет [1].

- Частичная аффилиация – автор является сотрудником университета и указал ссылку на несколько университетов [1].

Профиль автора a_i в информационной системе имеет следующий вид:

$$a_i = \langle K^i, P^i, Ind^i, W^i \rangle,$$

где K^i – множество ключевых слов i – го автора, P^i – множество публикаций, Ind^i – множество идентификаторов профилей в наукометрических базах данных, W^i – множество написаний автора на иностранном языке.

$$W^i = \{w^i{}_j\}_{j=1}^{t}$$

где t – количество уникальных иностранных написаний.

В качестве основного правила транслитерации была использована технология «OVIR of Russia regulations». В информационной системе университета предусмотрена возможность хранения различных вариантов транслитерации фамилии авторов, что позволяет использовать любые правила транслитерации и их комбинации. В связи с тем, что существуют различные методы транслитерации, не всегда возможно однозначно получить русскоязычное написание фамилии авторов. С учетом данного фактора возможно также и неоднозначное определение потенциальных авторов из базы физических лиц университета. Для обработки такой неоднозначности, необходима специализированная обработка данных [10]. В качестве обработки таких данных был разработан модуль анализа авторских коллективов публикаций авторских профилей Apr, наиболее схожих по написанию с a_i. Метод идентификации авторов заключается в определении потенциальных авторов Apt по написанию W авторов статьи p_{pr} с учетом научных коллективов и частоты их появления $f(a_{pt})$.

$$Apt = \left\{a_{pt}{}_j\right\}_{j=1}^{r},$$

где r – количество потенциальный сотрудников, подходящих написанию W.

В данной статье научные коллективы представлены следующим образом:

$$Akl^i = \langle Ca^i, Cw^i, St^i, Pt^i \rangle$$

где Ca^i – соавторы по публикациям автора a_i, Cw^i – сотрудники подразделений в котором работает или работал a_i, St^i – обучающиеся под руководством a_i, Pt^i – участники проектов, в которых участвует a_i.

На рисунке 3 представлены возможные варианты идентификации авторов. Рассмотрим пример, представленный на рисунке 3а, более детально. У публикации на английском языке указаны два автора: Dzerzhauskaya T.A., Varenikov D.A. Для того чтобы идентифицировать сотрудников, являющихся авторами данной публикации, необходимо по иностранному написанию фамилии, имени и отчеству найти в базе данных соответствующих сотрудников [2]. Для рассматриваемого примера были найдены следующие совпадения:

1. Автор 1 - Dzerzhauskaya T.A. Для данного автора были найдены следующие схожие написания:

- $w^1{}_1$ – Dzerzhauskaya T.A. Данное написание указано у двух пользователей:
  - a_{pt_1} – Дзержавская Т.А.
  - a_{pt_2} – Державская Т.А.
- $w^1{}_2$ – Dziarzhauskaya T.A. Данное написание определено на основании анализа иностранного написания фамилий авторов, хранящихся в системе у одного сотрудника:
  - a_{pt_3} – Дзиржавская Т.А.

2. Автор 2 - Varenikov D.A. Для данного автора было найдено одно написание:

- $w^2{}_1$ – Varenikov D.A.
  - a_{pt_4} – Варенников Д.А.

Таким образом, однозначно определить связь Автора 1 с пользователем информационной системы невозможно, в отличие от Автора 2, для которого была найдена только одна связь с a_{pt_3}. Для того чтобы определить Автора 1, используется анализ авторских коллективов. С помощью проведенного анализа удалось определить, что из потенциальных авторов a_{pt_1}, a_{pt_2}, a_{pt_3} только сотрудник a_{pt_2} участвовал в авторском коллективе с сотрудником a_{pt_4}.

Кроме того, возможен вариант неоднозначного определения соавтора после анализа авторских коллективов (см. рис. 3 б) и дополнительных сведений об авторах, в этом случае система оставляет данного автора нераспознанным и формирует подсказу для специалиста, который в дальнейшем будет обрабатывать публикацию. Чем больше авторов приведено в публикации и чем полнее они описаны, тем точнее происходит идентификация авторов на основе авторских коллективов (см. рис 3 в). На рисунке 3 г показан пример неоднозначного определения автора после транслитерации. В данном примере идентификация соавтора происходит только после анализа авторского коллектива и обработки специалистами публикации, на основании рекомендаций, представленных системой. Данный пример демонстрирует наполнение авторского профиля различными вариантами транслитерации его фамилии, что в дальнейшем позволяет идентифицировать его более точно [1].

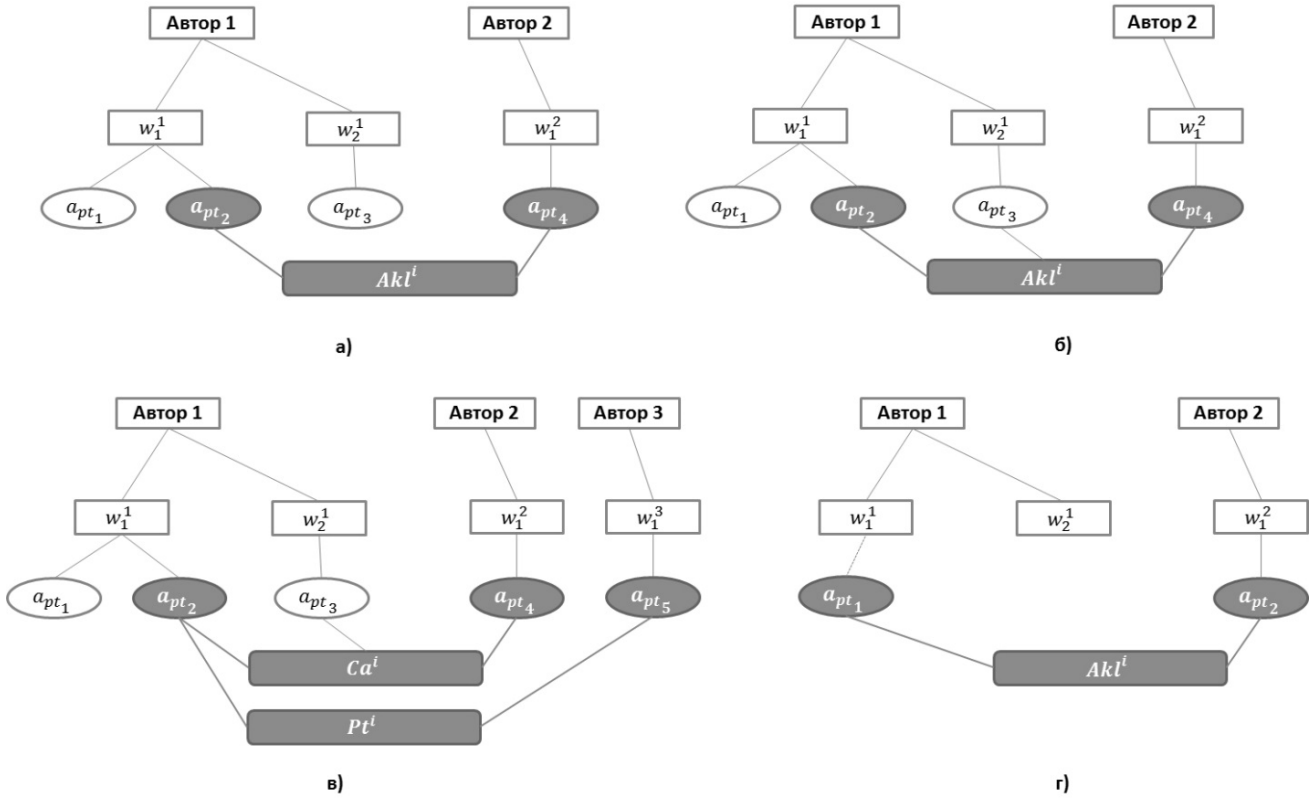

Рис. 3. Подход к идентификации авторов

Метод идентификации авторов, основанный на определении потенциальных авторов по написанию, с учетом научных коллективов и частоты их появления, позволил повысить качество определения и связи авторских профилей с наукометрическими базами данных и пользователями информационной системы. Рассмотренные подходы в дальнейшем будут применены к определению соответствия между пользователями информационной системы университета и их профилями в открытых научных Интернет-ресурсах [13].

Формирование профиля публикации

Информационная модель профиля публикации, представлена на Рис. 4. Одним из показателей профиля публикации являются ключевые слова K^s. Данный показатель важен при формировании рекомендаций и поиска публикаций.

$$K^s = \bigcup_{i=1}^{l} K^s{}_i,$$

где l – количество источников ключевых слов для периодического издания. Множество ключевых слов публикаций формируются на основе:

- Множества ключевых слов, указанных авторами, – $K^s{}_1$;
- Множества ключевых слов, полученных из наукометрических баз данных, – $K^s{}_2$. В наукометрических базах данных существует отдельное описание публикаций ключевыми словами и тематиками, соответствующим справочникам конкретной наукометрической базы;

- Множества ключевых слов периодического издания, к которому относится данная публикация, – $K^s{}_3$.

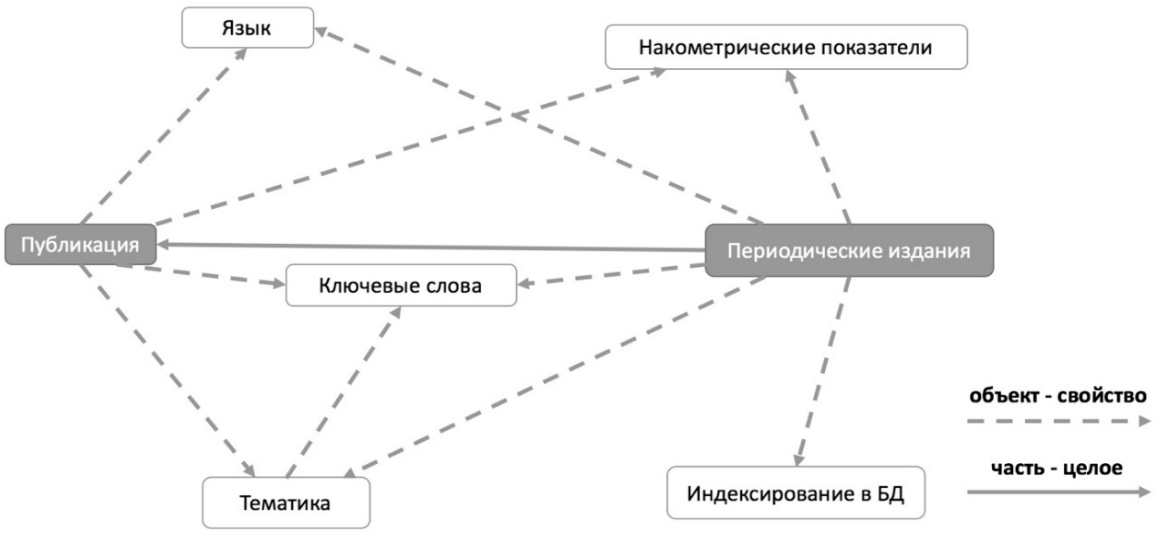

Рис. 4. Информационная модель профиля публикации

Формирования научных профилей конкурсов и научных мероприятий производится по схожей схеме и в статье не рассматриваются.

Заключение

В результате выполненной работы предложены подходы по автоматизации формирования научных профилей, которые позволили значительно расширить выборку, на основе которой в дальнейшем строятся рекомендации для пользователей информационной системы по выбору научного руководителя или обучающегося, научного мероприятия, грантов, публикаций и периодических изданий для публикации научных результатов. Полнота полученных данных позволила оптимизировать учет публикаций специалистами и, как следствие, повысить качество отчетных данных. Предложенные методы были реализованы в информационной системе управления университета.

Список литературы

1. Варенников Д.А., Муромцев Д.И., Шлей М.Д. Подходы автоматизации обработки данных наукометрических баз данных // Компьютерные инструменты в образовании. - 2015. - № 2. - С. 3-13.

2. Варенников Д.А., Шлей М.Д., Иванов В.В. Методы идентификации авторов при автоматизированной обработке информации о публикациях // Научно-образовательная информационная среда XXI века: Материалы IX Всероссийской научно-практической конференции, Петрозаводск, 23-25 сентября 2015 г. - 2015. - С. 36-39.

3. Ефимов М.Н., Шлей М.Д., Варенников Д.А. Метод определения рекомендаций для пользователей информационной системы на основе их научных интересов и активности // Научно-образовательная информационная среда XXI века. Материалы VIII Международной научно-практической конференции. Петрозаводск, 2014. - 2014. - С. 74-77.

4. Ефимов М.Н., Шлей М.Д., Варенников Д.А. Система определения научных интересов пользователей // Труды XXI Всероссийской научно-методической конференции "Телематика'2014". - 2014. - С. 87-88.

5. Казин Ф.А., Биккулов А.С., Зленко А.Н., Тойвонен Н.Р., Попова И.А., Шлей М.Д., Варенников Д.А. Система поддержки проектной деятельности в Университете ИТМО // Инновации. - 2014. - № 8(190). - С. 77-83.

6. Коляда А.С., Гогунскиий В.Д. Автоматизация извлечения информации из наукометрических баз данных // Управління розвитком складних систем. - 2013. - № 16. - С. 96 - 99.

7. Коэффициент Жаккара [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org/wiki/Коэффициент_Жаккара. - Загл. с экрана.

8. Мазов Н.А., Гуреев В.Н. Проблемы идентификации метаданных в наукометрических базах данных Web of Knowledge, Scopus и РИНЦ на примере профилей авторов // Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса: 19-я междунар. конф. «Крым 2012» (2-10 июня 2012 г., г. Судак): Труды конф. - М.: Изд-во ГПНТБ России, 2012. - С. 1-4.

9. Наукометрические базы данных [Электронный ресурс]. - Режим доступа: http://pspu.ru/university/biblioteka/prepodavatelam/indeksy-nauchnogo-citirovanija, свободный. – Загл. с экрана.

10. Пинжин А. Е. Применение вероятностного алгоритма соединения записей для исключения дублирования информации в корпоративной базе данных // Известия Томского политехнического университета. – 2006. – №7. – С. 111-116.

11. Попова И.А., Тойвонен Н.Р., Варенников Д.А. Система информационной поддержки проектной деятельности вуза // Информационные системы для научных исследований: Труды XV Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2012). - 2012. - С. 156-159.

12. Попова И.А., Громов Г.Ю. Подходы к созданию эффективной информационной системы управления университетом // Сборник материалов XX Всероссийской научно-методическая конференции. – Санкт-Петербург, 2013.

13. Семерханов И.А., Муромцев Д.И. Интеграция информационных систем на основе технологии связанных данных // Научно-технический вестник информационных технологий, механики и оптики. - 2013. - № 5 (87). - С. 123-127.

Рецензенты:

Арустамов С.А., д.т.н., профессор, профессор кафедры проектирования и безопасности компьютерных систем, Университет ИТМО, г. Санкт-Петербург;

Коробейников А.Г., д.т.н., профессор, заместитель директора по науке СПбФ ИЗМИ РАН, г. Санкт-Петербург.