УДК 004.6

И.В. Шостак, А.А. Лысенко

Национальный аэрокосмический университет им. Н. Е. Жуковского «ХАИ», Украина

МОДЕЛЬ СОЦИАЛЬНОГО ПРОФИЛЯ КАК БАЗА ПРЕДПОЧТЕНИЙ ПОЛЬЗОВАТЕЛЯ В ПРОЦЕССЕ СЕМАНТИЧЕСКОГО ПОИСКА ПО SEMANTIC WEB

Рассматривается постановка задачи определения модели пользователя как основы анализа предпочтений пользователя. Приведена структура профиля в формате свойств ресурса Semantic Web, а также формат хранения в RDF-документе, поддерживающем JSON стандарт. Получена ключевая компонента, которая может быть применена как составляющая социальной сети, направленной на семантический поиск.

Ключевые слова: *семантический поиск, социальный профиль, социальное индексирование документов, социальная сеть.*

Введение

С развитием сети Internet поисковым системам стала отводиться главенствующая роль. Благодаря их алгоритмам пользователи могли получить быстрый доступ к любой информации. Однако стремительное увеличение количества сайтов также начало порождать и проблему выбора. В современных условиях, чтобы найти релевантную информацию, которая соответствует критериям пользователя, необходимо потратить немало времени на обработку различных источников по интересующей его тематике. Это связано с наличием сайтов, содержимое которых зависит от информационных блоков, которые могут быть значимыми, так и незначимыми.

Таким образом, разработчики поисковых машин постоянно дорабатывают и разрабатывают различного рода методы поиска, обработки и представления информации. Тем самым идет развитие двух фундаментальных методов поиска – статического и семантического, - а в качестве мейнстрима все чаще стали применяться социальные сети.

Социальные сети позволяют обменивать информацией между участниками с близкими интересами, что в значительной мере облегчает поиск той или иной информации.

1. Обзор аналитической литературы

С учетом возрастающих объемов данных сети Internet на передний план выходит такая проблема, как автоматизация методов работы с большими объемами информации [1]. Эта задача решается при использовании новых стандартов Semantic Web.

Semantic Web – это ключевая составляющая концепции развития сети Интернет, суть которой заключается в реализации возможности машинной обработки информации, доступной во Всемирной паутине. Основной акцент концепции падает на работу с промежуточными данными, характеризующими свойства и содержание ресурсов Всемирной паутины. Основная идея Semantic Web – это надстройка над существующей сетью Интернет, которая призвана сделать размещенную в ней информацию более понятной для компьютеров [2].

Успешному развитию технологии Semantic Web способствовало использование универсальных идентификаторов ресурсов (Uniform Resource Identifier, сокр. URI) [2]. Стандартная схема использования таких идентификаторов в современном Интернете сводится к установке ссылок, ведущих на адресуемый объект, который можно загрузить и сохранить/просмотреть. Такими объектами могут быть:

- Web-страницы
- файлы произвольного содержания
- фрагменты Web-страниц
- неявное указание на обращение к реально существующим физическим ресурсам по протоколу, отличному от HyperText Transfer Protocol (сокр. HTTP).

Концепция семантической паутины расширяет это понятие, включая в него ресурсы, недоступные для скачивания. Адресуемыми с помощью URI ресурсами могут быть, например, конкретные люди, географические сущности, художественные артефакты и т.д.

Стоит отметить, что проблеме повышения эффективности работы всех типов поисковых систем [3] предоставлено множество работ [4-23], но, тем не менее, отсутствует универсальный подход к решению задач поиска и анализа релевантной Web-информации в сети Internet. В

представленных работах полученный результат выглядит в форме компромисса, при котором пользователю приходится мириться с некоторыми ограничениями или недостатками используемых им систем.

Таким образом, выделяют три вида систем: вербальные и классификационные, а также интеллектуальные агентов [3].

Вербальные поисковые системы [3] позволяют формировать базы данных разного объема, размер и диапазон которых зависят от конкретных системных задач. Эти поисковики характеризуются применением собственных языков запроса и алгоритмов ранжирования.

Классификационные поисковые системы или каталоги [3] основаны на принципе классификации информационных ресурсов (чем больше классов, тем более репрезентативна поисковая система).

Интеллектуальные агенты [3] – это программы, обычно дополняющие функции стандартных поисковых систем, и учитывающие, как правило, текущие предпочтения клиента.

С развитием поисковых систем усовершенствуются и методы поиска, анализа информации, а также формирования Web-страниц, характерные для выделенных типов систем [3]:

– статистические (которые автоматически индексируют пространство Internet путем перехода по ссылкам, найденные в процессе обработки Web-страниц);

– семантические (которые позволяют выявлять зависимости между различными Web-страницами и их тематикой);

– комбинированные (гибридные методы, заимствующие подходы решения задач обеих методов с целью возможности отслеживания новостной ленты в сети Internet любого содержания, а также отсеивание дублирующейся и зашумленной информации).

В статических методах задействованы интеллектуальные агенты [3], играющие роль поисковых роботов, занимающихся индексацией ссылок Web-страниц. Полученные данные группируются по ключевым словам (те, что несут наибольшую смысловую нагрузку в Web-документе), которые сокрыты в промежуточных данных, и сохраняются на сервере поисковой машины. Результатом поиска является выборка ссылок, ведущие к страницам, в которых имеются совпадения ключевых слов с пользовательским запросом. Каждый элемент данной выборки обладает своим приоритетом (или рейтингом), отражающим частоту совпадений. Однако данные методы не учитывают ни семантику слов, ни смысловую нагрузку запроса.

В качестве примера поисковой системы, построенной на таком методе, является Google. В ней учитывается взвешенный индекс цитируемости и ранг страницы [24].

Индекс цитирования – это количественный показатель, определяемый на основании популяризированных Web-документов в сети Internet.

Ранг страницы – это показатель, характеризующий значимость страницы. Ранг зависит от количества внешних ссылок на Web-документ.

Ранг можно определить при помощи следующей формулы:

$$PR(A) = (1-d) + d\sum_{i=1}^{n}\frac{PR(T_i)}{C(T_i)}, \qquad (1.1)$$

где A – анализируемая страница;

T - список страниц, указывающих на страницу A;

d - коэффициент затухания, равный, как правило, 0.85;

$PR(T_i)$ - вес страницы, указывающей на страницу A;

$C(T_i)$ - число ссылок со страницы T_i;

n - общее число страниц, указывающих на страницу A.

Поисковая система Google позволяет задавать такие ограничения на поиск как дата, язык, инициалы автора. Однако системы, построенные на таком типе методов, предоставляют не маленький список ссылок, указывающих на не всегда релевантные документы, поскольку в них отсутствует возможность выявления и исключения шумов при извлечении данных.

Помимо поисковых систем общего назначения, существуют еще и специализированные, работающие только с определенными типами данных [25]. Их преимущество заключается в использовании уникальных методов поиска. Например, существует случай, когда программный агент разделяет Web-страницу на содержательную и навигационную части путем использования специального алгоритма. Суть его заключается в выделении повторяющихся фрагментов страниц сайта.

Принцип работы алгоритма примерно следующий:

– на вход подается директория с файлами, соответствующим Web-страницам сайта;
– осуществляется анализ содержимого файлов и выделение повторяющихся фрагментов, формирующих навигационную часть;

– выделенные фрагменты помечаются специальными тегами, либо удаляются;
– оставшиеся фрагменты – это содержательная часть, – используемые при поиске документа в сети Internet.

Поисковые системы, основанные на семантическом типе методов, используют вспомогательные модули или eXtensible Markup Language (сокр. XML) - подобные файлы. Основной принцип работы данных методов заключается в предварительном формировании базы знаний, в которой находятся перечень характеристик, позволяющих определить лексический смысл исследуемого объекта.

Примерами использования данных методов в поисковых системах являются следующие поисковики:

– Kim-Semantic Annotation Platform [26];
– Similarity/Closeness-Based Resource Browser [27].

Kim-Semantic Annotation Platform - это поисковая система, позволяющая детализировано обрабатывать Web-страницы посредством формирования цепочек из Web-документов, описывающие тематические объекты. Такой подход обеспечивает пользователю доступ к любой информации об интересующем его объекте из любой просматриваемой им Web-страницы. Стоит отметить, что выбор возможен только из слов, которые выделены маркером в тексте.

Такое техническое решение достигается путем автоматической индексации Web-документов и построения их семантической аннотации. Семантические аннотации хранятся базе знаний в виде объектов (или ключевых слов). Другими словами, в Web-документе выделяются данные, соответствующие некоторым классам онтологии. Каждый класс может быть родительским, т.е. иметь производные классы, и имена этих классов должны соответствовать определенному термину. Таким образом, ключевой объект может соответствовать только одному классу.

В результате работы поисковой системы, все анализируемые Web-документы определяются к группам, которые соответствуют определенным предметным областям. Стоит отметить, что все данные в группе должны быть связаны между собой.

Особенностью Similarity/Closeness-Based Resource Browser является возможность схематически изображать семантическую близость Web-документов на основании смежности тематик. Тематики в свою очередь определяются при помощи онтологического инжиниринга [29]. Расчета такой близости основывается на применении эвклидова расстояния, либо показателя Normalized Google distance (сокр. NGD) [28].

Примером поисковой системы, использующей комбинированные методы, является InfoStream, которая позволяет формировать дайджесты, определять связь между рубриками, конструировать сюжетные цепочки др. [35]. Данная система обеспечивает интеграцию Web-документов путем применения средств отбора, обработки, хранения и доступа к данным. Сбор информации по Web-документам, а также ее структурирование по семантическим признакам выполняется в режиме реального времени.

Результатами работы системы являются аналитические отчеты, документы, либо RSS новости, предоставляемые пользователю по его электронным адресам.

В качестве сравниваемых характеристик методов рассматриваются их преимущества и недостатки, представленные в табл. 1.1.

Разработке поисковых систем посвящено множество работ различной тематики, в частности, проектирование и разработка таких систем, которые учитывают социальные связи между пользователями [31 - 34].

Проблема *поиска пользователями данных* может быть решена путем создания единого, доступного всем хранилища информации, содержащего документы по тематике, необходимой, либо интересующей, пользователей [31]. Каждый пользователь в свою очередь может ограничивать поиск по определенным атрибутам при помощи фильтра, т.е. определять список свойств, характеризующий искомую информацию. Этот список может быть расширен и другими пользователям. Таким образом, точность идентификации любого документа, хранимого в системе, значительно повышается.

Следует отметить, что использовать подобную систему могут те только пользователи, прошедшие регистрацию. Также общее количество пользователей определяет эффективность такой системы.

Следующая проблема – *мультиагентная организация обмена информацией между пользователями* – и пути ее решения представлены в работе [32]. При помощи программных агентов система формирует персональный пользовательский профиль, учитывающая активность пользователя за время его работы с системой. Каждый пользователь организует собственную рабочую среду, путем создания профиля в системе (идентифицирующую его как участника данной системы). В таком профиле приведены персональные данные, рабочие проекты, а также ссылки на сопутствующие Web-документы.

Система, анализируя данные профиля пользователя, формирует список ключевых слов, используемый в дальнейшем для поиска Web-документов по принципу совпадения ключевых слов со словами, встречаемых в текстовых документах. Результатом работы такой системы является определение рейтинга часто используемых источников информации для конкретного пользователя.

Относительно обмена информацией между пользователями осуществляется исходя из близости предметных областей их профилей.

Таблица 1.1. Характеристики методов

Тип методов	***Преимущества***	***Недостатки***
1	***2***	***3***
статистические	– поддержка различных типов данных Web-документов; – возможность конфигурирования ограничения поиска.	– отсутствие возможности выявления и исключения зашумленности Web-данных; – длинные результирующие списки ссылок, указывающие на Web-страницы; – нет гарантии, что найденная Web-страница является релевантной относительно пользовательского запроса.
семантические	– гарантия, что результаты поиска в большей степени удовлетворят требования пользователя; – возможность прогнозирования дальнейшие пожелания пользователя;	– ограничения семантического поиска смежных тематик различных Web-документов; – необходимость наличия заранее сформированных файлов, содержащих Web-страницы; – чувствительность к лексическим особенностям языка написания контента Web-страницы; – отсутствие возможности выявления и исключения зашумленности Web-данных; – невозможность конфигурирования ограничения поиска.
комбинированные	– поддержка различных типов данных Web-документов; – гарантия, что результаты поиска в большей степени удовлетворят требования пользователя; – возможность прогнозирования дальнейшие пожелания пользователя;	– низкий уровень проверки зашумленности Web-информации; – низкая скорость поиска; – невозможность конфигурирования ограничения поиска; – длинные результирующие списки ссылок, указывающие на Web-страницы.

Так как решение проблемы мультиагентного обмена информации между пользователями [32] приводит к тому, что система становится подобием социальной сети, то следующая проблема – это *обмен информацией между пользователями, которые объединены в социальную сеть в рамках организации* [33].

Подобные системы функционируют по принципу опроса пользователей, необходимого для оценки эффективности установления ограничений на обмен информации между различными отделами организации. Определено, что для пользователей важным является получение информации в сжатые сроки, при этом он полностью доверяет источнику информации. Алгоритм поиска ресурса основывается, как правило, на базе имеющихся в общем доступе ресурсах и опыта разработчика системы.

И последняя проблема, которая связана с определением значимых данных при использовании ресурса Digg [34]. Digg [35] – это динамическая Web-страница, которая позволяет пользователю оставлять ссылку на информацию и оценить другие источники информации, представленные

остальными пользователями, посредством голосования и комментирования.

В итоге, на главной странице ресурса приведены те источники информации, у которых высокий рейтинг. Разработчики системы оставляют пользователям право на объединение в социальные группы, в рамках которых рейтинг отмеченных ими ссылок выше, чем у прочих. Это способствует уменьшению влияния тех пользователей, которые неоправданно завышают рейтинг для источников информации.

Социальная сеть [36] – это Web-структура, которая состоит из агентов (индивидуальных или коллективных субъектов) и отношений между ними (сотрудничество, коммуникация, дружба).

Социальная сеть представима в виде графа (рис. 1.1): так в качестве вершин A_i выступают агенты, а в качестве ребер R_j – взаимосвязь агентов. Так как сеть – граф, то изменение состояния одной вершины приведет к неминуемому изменению остальных. Учитывая специфику задачи, можно сказать, что информация, доступная одному агенту, доступна всем агентам сети.

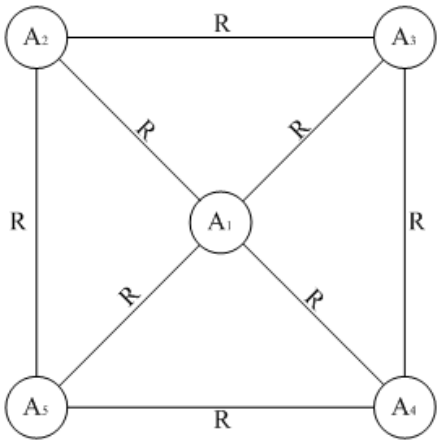

Рисунок 1.1. Формальное представление социальной сети в виде графа $G(A,R)$

Активное развитие социальных сетей обусловливается возможностью мгновенного обмена информацией между участниками сети, что в свою очередь обеспечивает возможность поиска информации на основании содержимого документов.

Существует критерий поиска по предрасположенности некоторой группы людей к интересующей их информации [37]– *социальный индекс* (критерий социальной оценки документа), который позволяет оценивать значимость документа с учетом актуальных интересов участника сети. Другими словами, документы могут быть проиндексированы не по ключевым словам, а по принадлежности к группе людей.

Социальный индекс документа для одного интереса участника сети определяется следующей формулой [37]:

$$SI = \sqrt{\frac{1}{2}(MU_t - MU)^2 + \frac{1}{2}(CU_t - CU)^2}\ , \quad (1.2)$$

где MU_t - оценка значимости шаблона документа;

MU - средняя оценка значимости документа среди участников со схожими интересами в диапазоне $[0;10]$;

CU_t - соотношение общего числа посещений шаблона к числу посещений участников с одним исследуемым интересом;

CU - соотношение общего числа документа к числу посещений участников с одним исследуемым интересом в диапазоне $[0;1]$.

В основе формулы (1.2), лежит определение евклидова расстояния с двумя независимыми друг от друга параметрами и весовыми коэффициентами (заданные относительно результатов экспертных оценок).

Принцип применения метода, заключенный на использование рассматриваемого критерия, заключается в выполнении следующих этапов:

- получение/определение списка актуальных значений CU и MU для Web-документа, которые задаются другими участниками сети;
- выборка определенных значений CU и MU, учитывающие некоторые интересы определенной группы людей;
- расчет социального индекса документа;
- если происходит повторный просмотр документа, необходимо провести перерасчет значения CU_t по всему списку значений с учетом текущего интереса пользователя;
- если происходит оценка документа, то – значение MU_t;
- если оценка документа производится некоторой группой людей с различными интересами, то социальный индекс рассчитывается для каждого участника сети по отдельности;
- тогда итоговое значение социального индекса определяется как сумма всех значений SI конкретных документов, которые соответствуют группе интересов текущего участника.

Применяя социальное индексирование как критерий поиска релевантной информации, поисковая система будет формировать список

ссылок на Web-документы в порядке соответствия интересам пользователя.

Актуальной задачей является способствование развития семантического поиска на основе социальных сетей путем изучения и введения в практику концепций Semantic Web.

2. Социальный профиль как модель пользователя

На сегодняшний день широкое распространение в области поиска и обменом информации получили социальные сети, объединяющие между собой пользователей WWW. Процесс объединения заключается в поиске и формировании групп пользователей с близкими интересами. Эти интересы указываются каждым пользователем в его собственном профиле, называемом социальным профилем, в котором, помимо прочего, отражены такие атрибуты как инициалы, возраст, место жительства, контактная информация и список групп, на которые подписан пользователь. В социальном профиле также отражены понравившиеся пользователю документы, либо те документы, с которыми пользователь желает ознакомиться в будущем, в формате URI.

Социальные сети при поддержке одной из ключевой функциональности - поиске информации – выполняют просмотр документов других пользователей с близкими интересами к пользователю, инициировавшему запрос на поиск. Другими словами, сеть, в своей работе, активно применяет наборы социальных профилей.

Таким образом, социальный профиль пользователя играет ключевую роль в сети, обеспечивая эффективность поиска информации всеми пользователями сети. Основываясь на Semantic Web [39], социальный профиль представим в качестве ресурса, содержащего свойства пользователя. Следует отметить, что применение Semantic Web поддерживает логический вывод [38], т.е. это поспособствует поддержке функциональности предложения просмотра документов на основании интересов пользователя, которые автоматически дополняются в процессе серфинга по сети.

2.1 Структура профиля

Каждый ресурс Semantic Web может иметь некоторый набор свойств. В данном случае, эти свойства, характеризующие пользователя. Эти свойства хранятся в RDF-документе. Для обеспечения эффективной работы сети, в таблице 2.1 приведен набор необходимых свойств.

Стоит отметить, что свойства являются также ресурсами. Для каждого ресурса существует категория, определяемая посредством RDF-классов, также являющиеся ресурсами (например: документ [foaf:Document], список [rdf:List], тип [rdf:type]).

Для задания свойства социального профиля применяется триплет (или высказывание о ресурсе), использующие язык утверждений RDF. Триплет состоит из трех частей [38]:

– субъект – описываемый ресурс;
– предикат – свойство ресурса;
– объект – значение свойства.

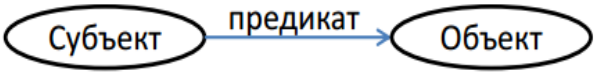

Рисунок 2.1. Графическое представление триплета

Каждый ресурс идентифицируется при помощи URI или не идентифицируется (тогда такой ресурс называется анонимным), и в случае значений свойств ресурса, - при помощи литералов (строка символов в кодировке Unicode).

Таблица 2.1. Свойства ресурса “Социальный профиль”

№	*Наименование*	*Категория*	*Назначение*
1	фамилия	rdf:type	отражает подпись пользователя в сети
2	имя	rdf:type	
3	возраст	rdf:type	определяет возраст пользователя
4	местонахождение_ по_государству	rdf:type	определяет местонахождение пользователя в контексте государства
5	местонахождение_ по_региону	rdf:type	определяет местонахождение пользователя в контексте региона государства
6	e-mail	rdf:type	определяет электронный ящик пользователя
7	телефон	rdf:type	определяет номер контактного телефона пользователя
8	список интересов	rdf:List	отражает предпочтения пользователя
9	список групп	rdf:List	определяет ссылки на группы, на которые подписан пользователь
10	список документов	rdf:List	определяет ссылки на документы, которые отметил пользователь

Структура URI имеет следующий вид [40]:

URI = [схема ":"] иерархическая_часть ["?" запрос] ["#" фрагмент]

где *схема* - схема обращения к ресурсу (часто указывает на сетевой протокол), например HTTP, FTP, FILE;

иерархическая часть - содержит данные, обычно организованные в иерархической форме, которые, совместно с данными в неиерархическом компоненте запрос, служат для идентификации ресурса в пределах видимости URI-схемы. Обычно часть содержит путь к ресурсу (и, возможно, перед ним, адрес сервера, на котором тот располагается) или идентификатор ресурса (в случае URN);

запрос - содержит данные.в произвольной форме, необходимые для детальной идентификации ресурса в пределах видимости URI-схемы;

фрагмент [41] - тоже необязательный компонент и позволяет косвенно идентифицировать вторичный ресурс посредством ссылки на первичный и указанием дополнительной информации. Вторичный идентифицируемый ресурс может быть некоторой частью или подмножеством первичного, некоторым его представлением или другим ресурсом, определённым или описанным таким ресурсом.

Часть идентификатора URI без схемы обращения к ресурсу часто называется «ссылкой URI».

В конечном счете, триплеты из RDF-документа объединяются в RDF-граф [45], представленный на рисунке 2.2. Все RDF-графы объединяются в единый граф Giant Global Graph (GGG). Giant Global Graph объединяет все связанные данные. Такая концепция является шагом в развитии WWW.

Список (или коллекция) подобен списку в языке Lisp. Для ее создания создается ресурс относящийся к классу rdf:list и имеющий два свойства:

- rdf:first (голова) — первый элемент списка;
- rdf:rest (хвост) — ссылка на подсписок, содержащий оставшиеся элементы.

Подсписок также содержит голову и хвост. Хвост подсписка, содержащий последний элемент указывает на встроенный ресурс rdf:nil.

2.2 Способы хранения RDF-графа

Существует два подхода реализации хранения RDF-графов [38]:

- Triplestore и точка доступа (аналогичен реляционной базе данных).
- Текстовый файл в одном из специальных форматов таблицы 2.2 (посредством сериализации).

Таблица 2.2. Форматы сериализации RDF-графа

№	*Формат*	*Описание*
1	RDF/XML	стандартный формат на базе XML.
2	N-Triples	состоит из простого перечисления триплетов.
3	Turtle	является расширением N-Triples и позволяет записывать триплеты в более понятном и компактном виде.
4	JSON-LD	формат на базе JSON.
5	RDFa и Microdata	формат RDF-разметки HTML-страниц.

Предлагается рассматривать формат RDF/JSON, т.к. в последнее время он более популярен за счет простого механизма структурного анализа во время проектирования систем.

2.3 Формат сериализации социального профиля

Формат RDF/JSON представляет наборы триплетов графа в виде последовательностей вложенных структур данных. Каждый уникальный субъект в наборе представим как ключ в JSON объекте (массиве, словаре, хэш-таблице). Значением каждого ключа есть объект, значение ключей которого являются URI свойств, ассоциирующихся с каждым субъектом. Значением каждого ключа свойства есть массив объектов, который предоставляет значения для каждого свойства.

В общем случае, триплет (субъект *S*, предикат *P*, объект *O*) представим в виде следующей структуры [42]:

{ "S" : { "P" : [O] } }

Объект *O* в триплете имеет вид объекта JSON со следующими ключами:

type – обязательный; может принимать одно из значений: “uri”, “literal”, “bnode”;

value – обязательный; описывает лексическое значение объекта;

lang – необязательный; определяет язык литералов ключа value;

datatype – необязательный; содержит URI на тип данных.

Ниже приведен шаблон RDF/JSON спецификации в формате JSON схемы:

```
{  "version":"0.3.0",
   "id":"RDF-JSON",
   "description":"RDF JSON definition",
   "type":"object",
   "properties":{     },
   "additionalProperties":{ "type":"object",
      "description":"subject (root object)",
      "optional":"true",
      "properties":{        },
```

```
"additionalProperties":{ "type":"array",
  "description":"predicate (subject object)",
  "optional":"true",
  "items":{ "type":"object",
    "description":"object (value array)",
    "properties":{
      "description":"content (value object)",
      "type":{ "type":"string",
        "enum":["uri","bnode","literal"] },
      "value":{ "type":"string" },
      "lang":{ "optional":true,
        "description":"See ftp://ftp.isi.edu/
        in-notes/bcp/bcp47.txt",
        "type":"string" }
    }
  }
}
}
```

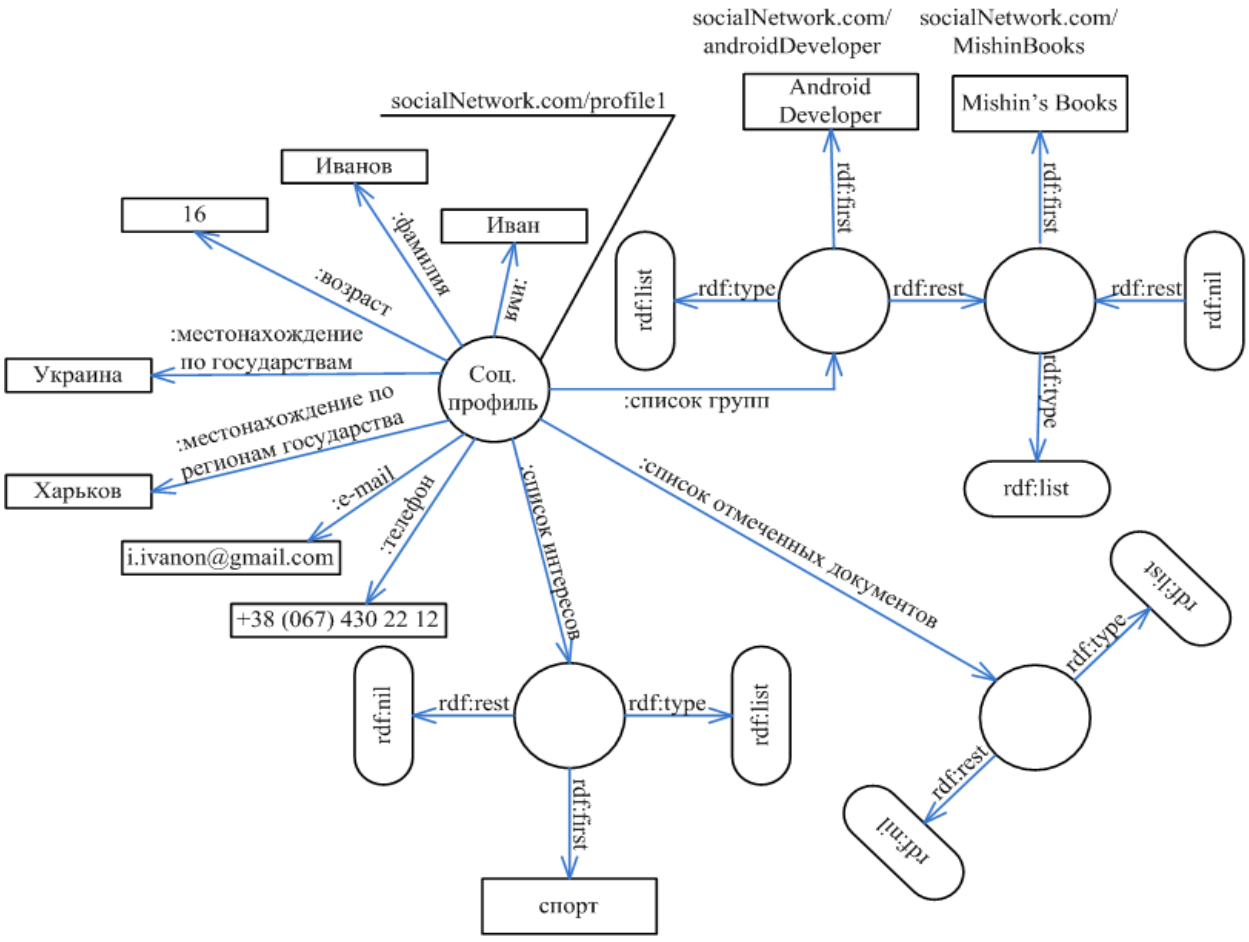

Рисунок 2.2. Графическое представление RDF-документа, характеризующего социальный профиль

В соответствии с шаблоном, представление социального профиля в формате RDF/JSON будет иметь вид:

```
{
"http://socialNetwork.com/profile1" : {
    "http://www.w3.org/1999/02/22-rdf-syntax-
     ns#/type" : {
          "value" : "http://xmlns.com/foaf/0.1/
           Person",
          "type" : "uri" },
    "http://xmlns.com/foaf/0.1/Person/firstName" : {
          "value" : "Иван", "type" : "literal" },
    "http://xmlns.com/foaf/0.1/Person/secondName" : {
          "value" : "Иванов", "type" : "literal" },
    "/groups" : [
          "http://socialNetwork.com/
           androidDeveloper",
          "http://socialNetwork.com/MishinsBooks" ]
}
}
```

2.4 Программный интерфейс Jena RDF API как инструмент формирования RDF-графов

Jena (или Apache Jena) [43] – это бесплатная и с открытым исходным Java кодом программная платформа, назначение которой поддержка построения семантической сети [43, 44] и связанных данных приложений [43, 45]. Программная платформа состоит из различных программных интерфейсов (см. рис. 2.3), взаимодействующих друг с другом в процессе обработки RDF-файлов.

Однако в данной работе внимание направлено только на интерфейс представления графа (подобного рис. 2.2) в RDF-файле – Apache Jena RDF API. Программный интерфейс содержит классы для представления графов, ресурсов, свойств и литералов (основываясь на терминологии RDF).

Таким образом, в таблице 2.3 представлена ассоциация терминов по RDF и классов Jena, позволяющих оперировать терминами при помощи экземпляров этих классов соответственно.

Применение объектов Resource и Model в собственных разработках сводится к созданию экземпляров классов при помощи методов createResource(…) и createDefaultModel() соответственно, поддерживая паттерн проектирования “Фабрика”.

Что касается литералов, то объекты этого класса могут быть использованы как при назначении свойств ресурсу, так и при работе с самим ресурсом, например, созданием ресурса, обладающим определенным URI.

Свойствами ресурсов может быть как одна запись, так и последовательность (или коллекция) записей. Во втором случае применяются контейнеры различных видов (см. рис. 2.4), элементы которых могут быть как литералы, так и ресурсы.

Результаты

Раскрыто значение социального профиля пользователя как вспомогательный компонент семантического поиска, определяющий интересы пользователя.

Сформирована структура профиля пользователя, содержащая поля, которые могут быть полезны в процессе поиска.

Приведен формат содержимого RDF-документа, хранящего свойства социального профиля в одном из сериализуемых форматов документов такого типа – RDF/JSON.

Выводы

Для поддержки развития семантического поиска на базе социальных сетей следует искать попытки к поиску технических решений и, как вариант, использование модели пользователя как источника анализа интересов пользователя и хранилище цифровых документов в реализуемом прототипе сети, направленной на семантический поиск.

Таблица 2.3. Ассоциация терминов по RDF к классам программного интерфейса Apache Jena RDF API

№	Термины по RDF	Реализация в Jena	Примечания
1	ресурсы	Resource	-
2	свойства	Properties	-
3	литералы	Literal	-
4	граф	Model	В Jena граф - это модель.

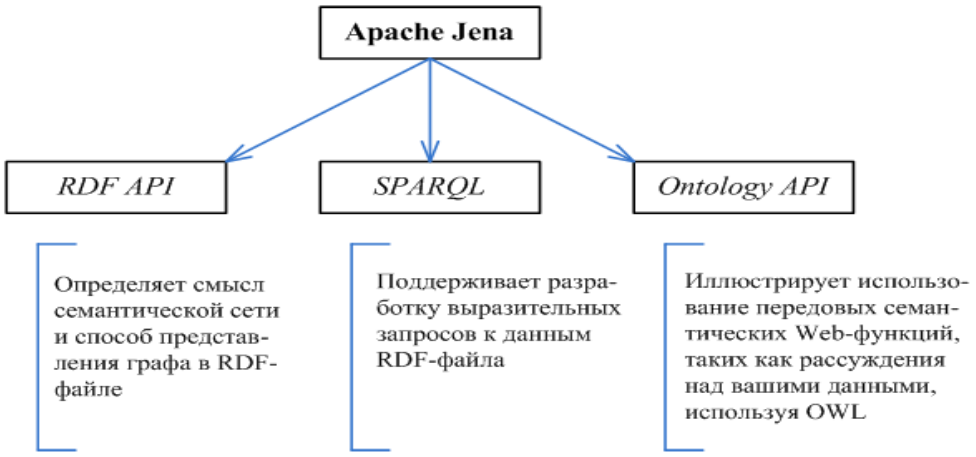

Рисунок 2.3. Состав программной платформы Jena

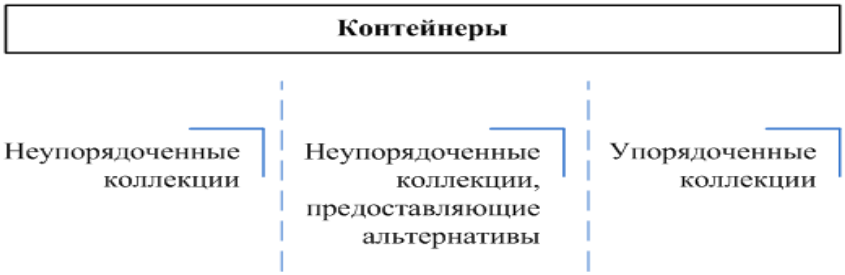

Рисунок 2.4. Разновидности контейнеров в Apache Jena

Литература

1. *Браславский П. Автоматическое реферирование веб-документов с учетом запроса [Текст] / П. Браславский // Интернет-математика 2005. Автоматическая обработка веб-данных. – М. : "Яndex", 2005. – С. 485-501.*
2. *Дементьев А.О. Семантическая паутина или будущее сети интернет [Текст] / А.О. Дементьев // Журнал: «Вопросы современной науки и практики Университет имени В.И. Вернадского». Симферополь: ТГТУ,2008. – Том 2, Вып. 4(14). – С. 59-64.*
3. *Захаров. В. Информационные системы (документальный поиск): [Текст] / В.Захаров. – СПб.: Изд-во СПбГУ, 2002. – 188 с.*
4. *Telnet protocol specification [Электронный ресурс] / J. Postel, J. Reynolds – Электрон. текст. дан. – Режим доступа: www/ URL: http://tools.ietf.org/html/rfc854 – 03.12.2011 г. – Загл. с экрана.*
5. *FAQ по поисковой системе WAIS [Электронный ресурс] / Д. Руденко – Электрон. текст. дан. – Режим доступа: www/ URL: http://faqs.org.ru/internet/wais.html – 10.12.2011 г. – Загл. с экрана.*
6. *Rashid H. From card Catalogue to Web OPACs [Текст] / H. Rashid, A. A. Mehtab. // DESIDOC Bulletin of Information Technology. – 2006. – Vol. 26 (2). – P. 41-47*
7. *Kelly M. Information Retrieval (Z39.50): Application Server Definition and Protocol Specification [Текст] / M. Kelly. – USA, Maryland: "NISO", 2002. – 276 p.*
8. *Valasqyez J. Adaptive Web Site. A Knowledge Extraction from Data Approach [Текст] / J. Valasqyez, V. Palade. – Amsterdam, Berlin, Oxford, Tokyo, Washington, Dc: IOS Press, 2008. – 272 p.*
9. *Томас М. Секреты программирования для Internet на Java [Текст] / М. Томас, П. Патлер А. Хадсон и др. – Спб.: Питер, 2002. – 390 с.*
10. *Вильямсон Х. Универсальный Dynamic HTML [Текст] / Х. Вильямсон. – Спб.: Питер, 2001. – 304 с.*
11. *Sadat H. On The Evaluation of AdaptiveWeb Systems [Текст] / H. Sadat, A.A. Ghorbani // The Second International Workshop on Web-based Support Systems in conjunction with AI. – Beijing, China, 2004. – P. 127-136.*
12. *Perkowitz M. Adaptive web sites: an ai challenge [Текст] / M. Perkowitz, O. Etzioni // The 15th International Joint Conference on Artificial Intelligence. – 1997. – P. 16-23.*
13. *Perkowitz M. Towards adaptive Web sites: Conceptual framework and case study [Текст] / M. Perkowitz, O. Etzioni // Artificial Intelligence. – USA, Washington, 2000. – Vol. 118 – P. 245-275.*
14. *Kamdar T. On creating adaptive web servers using weblog mining [Текст] / T. Kamdar, A. Joshi // Technical Report, TR-CS-00-05, CS Department, UMBC. – 2000. – 18 p.*
15. *Pazzani M. Adaptive Web Site Agents [Текст] / M. Pazzani, D. Billsus // Journal of Agents and Multiagent systems. – 2002. – Vol. 5(2). – P. 205-218.*
16. *Gibson J. Adaptive Web-page content identification [Текст] / J. Gibson, B. Wellner, S . Lubar // The 9th annual ACM international workshop on Web information and data management. – USA, New York, 2007. – P. 105-112.*
17. *Balabanovic M. An adaptive web-page recommendation service [Текст] / M. Balabanovic // The first International conference on autonomous agent. – USA, New York, 1997. – P. 378-385.*
18. *Касьянова Е.В. Адаптивные методы и средства поддержки дистанционного обучения программированию [Текст] / Е.В. Касьянова. – Институт систем информатики имени А.П. Ершова: Новосибирск, 2007. – 171 с.*
19. *Асеев Г.Г. Методы Интеллектуального анализа данных в электронных хранилищах [Текст] / Г.Г. Асеев // Научно-технический журнал «Бионика интеллекта». – 2009. – Вып.1(70). – С. 28-33.*
20. *Ланде Д.В. InfoStream. Мониторинг новостей из Интернет: технология, система, сервис [Текст] / Д.В. Ланде, А.Н. Григорьев, С.А. Бороденков и др. – К.: ООО "Старт 98", 2007. – 40 с.*
21. *Popov B. Semantic annotation platform [Текст] / B. Popov, A. Kiryakov, D. Manov // Natural language engineering. – Cambridge, 2004. – Vol. 10, No. 3-4. – P. 375-392.*
22. *Krhriyenko O. Similarity/closeness-based resource browser [Текст] / O. Krhriyenko. V. Terziyan // Visualization, imaging, and image processing. – Cambridge, 2009. – P. 184-191.*
23. *Naumenko A. Service matching in agent system [Текст] / A. Naumenko, S. Nikitin, V. Terziyan // Applied intelligent. – MA, USA, 2006. – Vol. 25/2. – P. 223-237.*
24. *Яковлев А. А. Раскрутка и продвижения сайтов: основы, секреты трюки [Текст] / А. А. Яковлев. – СПб.: Бхб-Петербург, 2007. – 336 с.*
25. *Агеев М. С. Извлечение значимой информации из web-страниц для задач информационного поиска [Текст] / М. С. Агеев, И. В. Вершинников, Б. В. Добров // Интернет-математика 2005. Автоматическая обработка веб-данных. – М.: "Яndex", 2005. – С. 283-301.*
26. *Popov B. Semantic annotation platform [Текст] / B. Popov, A. Kiryakov, D. Manov // Natural language engineering. – Cambridge, 2004. – Vol. 10, No. 3-4. – P. 375-392.*
27. *Krhriyenko O. Similarity/closeness-based resource browser [Текст] / O. Krhriyenko. V. Terziyan // Visualization, imaging, and image processing. – Cambridge, 2009. – P. 184-191.*

28. Thuraisingham Bh. Building trustworthy Semantic Web [Текст] / Bh. Thuraisingham. – NY, USA: Auerbach Publication, 2008. – 402 p.
29. Cilibrasi R. The google similarity distance [Текст] / R. Cilibrasi, P. Vitanyi // IEEE Transactions on knowledge and data engineering. – NJ, USA, 2007. – Vol. 19/3. – P. 370-383.
30. Ланде Д.В. InfoStream. Мониторинг новостей из Интернет: технология, система, сервис [Текст] / Д.В. Ланде, А.Н. Григорьев, С.А. Бороденков и др. – К.: ООО "Старт 98", 2007. – 40 с.
31. Raje R. R. On On designing and implementing a collaborative system using the distributed-object model of Java RMI [Текст] / R. R. Raje, S. Mukhopadnyay, M. Boyles and others // Progress in computer research. – Nova Science Publishers, Inc. Commack, NY, USA, 2001. – P. 123-134.
32. Vivacqua A. Profiling and matchmaking strategies in support of opportunistic collaboration [Текст] / A. Vivacqua, M. Moreno, J. Souz // Lecture notes in computer science. – Springer-Verlag, Berlin Heidelberg, 2003. – P. 162-177.
33. McDonald D. W. Recommending collaboration with social networks: A comparative evaluation [Текст] / D. W. McDonald // Proceedings of the SIGNCHI conference on Human factors in computing systems. – 2003. – P. 593-600.
34. Lerman K. Social networks and Social information filtering on Digg [Текст] / K. Lerman // Computing Research Repository – 2006. – 8 p.
35. Digg – What the Internet is talking about right now [Электронный ресурс] / – Режим доступа: www/ URL: http://digg.com/ – 27.03.2012 г. Загл. с экрана.
36. Губанов Д.А. Социальные сети: модели информационного влияния, управления и противоборства [Текст] / Д.А. Губанов, Д.А. Новиков, А.Г. Чартишвили. – М.: Физматлит, 2010 – 228 с.
37. Почанский О.М. Методы синтеза адаптивных web-страниц на основе интеллектуального анализа информационных ресурсов сети internet: Дис. ... канд. техн. наук: 05.13.23. – Х., 2012. – 168с.
38. Обучающий материал "Технологии Semantic Web" [Электронный ресурс] / - Режим доступа: http://habrahabr.ru/company/itis/blog/258405.
39. Semantic Web [Электронныйресурс] / - Режим доступа: http://semanticweb.org.
40. URI [Электронныйресурс] / - Режим доступа: https://ru.wikipedia.org/wiki/URI.
41. Uniform Resource Identifier (URI): Generic Syntax [Электронныйресурс] / - Режим доступа: https://tools.ietf.org/html/rfc3986.
42. Jena documentation overview. RDF JSON [Электронный ресурс] / - Режим доступа: http://jena.apache.org /documentation/io/rdf-json.html.
43. Getting started with Apache Jena [Электронный ресурс] / - Режим доступа: http://jena.apache.org/getting_started/ index.html
44. Wikipedia: Semantic Web [Электронный ресурс] / - Режим доступа: https://en.wikipedia.org/wiki/Semantic_Web.
45. Linked Data [Электронный ресурс] / - Режим доступа: http://linkeddata.org.

References

1. Braslavsky P (2005). Web documents automatic referencing based on the query. Internet Mathematics 2005. Automated processing of Web data, 485-501.
2. Dementyev A.O (2008). Semantic Web or the Internet future. Magazine "Modern science and practice questions of the University named after V.I. Vernadsky", 4 (14), 59-64.
3. Zakharov. V (2002). Information systems (document search). Publishing house of St. Petersburg State University, 188.
4. J. Postel, J. Reynolds. Telnet protocol specification. Retrieved from http://tools.ietf.org/html/rfc854.
5. D. Rudenko. FAQ of search system WAIS Retrieved from http://faqs.org.ru/internet/wais.html.
6. H. Rashid, A. A. Mehtab (2006). From card Catalogue to Web OPACs. DESIDOC Bulletin of Information Technology, 26 (2), 41-47.
7. Kelly M (2002). Information Retrieval (Z39.50): Application Server Definition and Protocol Specification. USA, Maryland: "NISO", 276 p.
8. J. Valasqyez, V. Palade (2008). Adaptive Web Site. A Knowledge Extraction from Data Approach. Amsterdam, Berlin, Oxford, Tokyo, Washington, Dc: IOS Press, 272 p.
9. M. Thomas P. Patler A. Hudson (2002). Secrets of Internet programming in the Java. SPb.: Peter, 390.
10. Williamson H (2001). Universal Dynamic HTML. SPb .: Peter, 304.
11. H. Sadat, A.A. Ghorbani. On The Evaluation of Adaptive Web Systems. The Second International Workshop on Web-based Support Systems in conjunction with AI, 127-136.
12. M. Perkowitz, O. Etzioni (1997). Adaptive web sites: an ai challenge. The 15th International Joint Conference on Artificial Intelligence, 16-23.
13. M. Perkowitz, O. Etzioni (2000). Towards adaptive Web sites: Conceptual framework and case study. Artificial Intelligence, 118, 245-275.
14. T. Kamdar, A. Joshi (2000). On creating adaptive web servers using weblog mining . Technical Report, TR-CS-00-05, 18.
15. M. Pazzani, D. Billsus (2002). Adaptive Web Site Agents. Journal of Agents and Multiagent systems, 5(2), 205-218.
16. J. Gibson, B. Wellner, S . Lubar (2007). Adaptive Web-page content identification. The 9th annual ACM international workshop on Web information and data management, 105-112.
17. Balabanovic M (1997). An adaptive web-page recommendation service. The first International conference on autonomous agent, 378-385.
18. Kasyanov E.V (2007). Adaptive methods and resources to support distance learning programming. Institute of Informatics Systems named after A.P. Yershov, 171 p.
19. G.G. Aseev (2009). Data mining techniques in electronic storage. Scientific and Technical Journal "Bionics intelligence", 1 (70), 28-33.
20. D.V. Lande, A.N. Grigoriev, S.A. Borodenko (2007). Monitoring of news from the Internet: technology, system service. Profile "Start 98", 40.
21. B. Popov, A. Kiryakov, D. Manov (2004). Semantic annotation platform. Natural language engineering, Vol.10, 3-4, 375-392.

22. O. Krhriyenko. V. Terziyan (2009). Similarity/closeness-based resource browser. Visualization, imaging, and image processing, 184-191.
23. A. Naumenko, S. Nikitin, V. Terziyan (2006). Service matching in agent system. Applied intelligent, 25/2, 223-237.
24. Yakovlev A.A (2007). Website promotion: bases, secret tricks. SPb.: Bhb Petersburg, 336.
25. M.S. Ageev, I.V. Vershinnikov, B.V. Dobrov (2005). Extracting meaningful information from web-pages for information retrieval tasks. Internet Mathematics 2005. Automated processing of Web data, 283-301.
26. V. Popov, A. Kiryakov, D. Manov (2004). Semantic annotation platform. Natural language engineering, Vol. 10, No. 3-4, 375-392.
27. O. Krhriyenko. V. Terziyan (2009). Similarity/closeness-based resource browser. Visualization, imaging, and image processing, 184-191.
28. Thuraisingham Bh (2008). Building trustworthy Semantic Web. Auerbach Publication, 402.
29. R. Cilibrasi, P. Vitanyi (2007). The Google similarity distance. IEEE Transactions on knowledge and data engineering, 19/3, 370-383.
30. D.V. Lande, A.N. Grigoriev, S.A. Borodenko (2007). Monitoring of news from the Internet: technology, system service, Profile "Start 98",40.
31. R. R. Raje, S. Mukhopadnyay, M. Boyles and others (2001). On designing and implementing a collaborative system using the distributed-object model of Java RMI. Progress in computer research, 123-134.
32. A. Vivacqua, M. Moreno, J. Souz (2003). Profiling and matchmaking strategies in support of opportunistic collaboration. Lecture notes in computer science, 162-177.
33. McDonald D. W. (2003). Recommending collaboration with social networks: A comparative evaluation. Proceedings of the SIGNCHI conference on Human factors in computing systems, 593-600.
34. Lerman K. (2006). Social networks and Social information filtering on Digg . Computing Research Repository, 8.
35. Digg – What the Internet is talking about right now Retrieved from http://digg.com.
36. D.A. Gubanov, D.A. Novikov, A.G. Chartishvili (2010). Social Networks: Models of information influence, control and confrontation. FIZMATLIT, 228.
37. Pochansky O.M. (2012). Methods of synthesis of adaptive web-pages on the basis of mining information resource network internet, 168.
38. Training material "Technologies Semantic Web". Retrieved from http://habrahabr.ru/company/itis/blog/258405.
39. Semantic Web. Retrieved from http://semanticweb.org.
40. URI. Retrieved from: https://ru.wikipedia.org/wiki/URI.
41. Uniform Resource Identifier (URI): Generic Syntax. Retrieved from https://tools.ietf.org/html/rfc3986.
42. Jena documentation overview. RDF JSON. Retrieved from http://jena.apache.org/documentation/io/rdf-json.html.
43. Getting started with Apache Jena [Электронный ресурс]. Retrieved from http://jena.apache.org/getting_started/index.html
44. Wikipedia: Semantic Web. Retrieved from https://en.wikipedia.org/wiki/Semantic_Web.
45. Linked Data. Retrieved from http://linkeddata.org.

Автор: ШОСТАК Игорь Владимирович
доктор технических наук, профессор, профессор кафедры Программная инженерия
Национальный аэрокосмический университет им. Н. Е. Жуковского «ХАИ», Украина
E-mail – iv.shostak@gmail.com

Автор: ЛЫСЕНКО Александр Александрович
магистрант кафедры Программная инженерия
Национальный аэрокосмический университет им. Н. Е. Жуковского «ХАИ», Украина
E-mail – alysenko94@gmail.com

МОДЕЛЬ СОЦІАЛЬНОГО ПРОФІЛЮ ЯК БАЗА ВПОДОБАННМ КОРИСТУВАЧІВ У ПРОЦЕСІ СЕМАНТИЧНОГО ПОШУКУ ПО SEMANTIC WEB

І.В. Шостак, А.А. Лисенко

Розглядається постановка задачі визначення моделі користувача як основи аналізу уподобань користувача. Наведено структуру профілю у форматі властивостей ресурсу Semantic Web, а також формат зберігання в RDF-документі, який підтримує JSON стандарт. Отримано ключову компонента, яка може бути застосована як складова соціальної мережі, спрямованої на семантичний пошук.

***Ключові слова**: семантичний пошук, соціальний профіль, соціальне індексування документів, соціальна мережа.*

THE SOCIAL PROFILE MODEL AS BASE OF USERS INTERESTS IN SEMANTIC SEARCH ACCORDING TO THE SEMANTIC WEB

I.V. Shostak, A.A. Lysenko

The problem formulation of the user model used during semantic search as base of interests is reviewed. The profile structure with properties of Semantic Web is shown and the profile is extension of existing type Semantic Web resources which holds data about user's interests in the Internet. This researchable model will be able use in system with semantic search functionality within such model representation as file based on RDF/JSON. The model representation supports by serialization mechanism of chosen Jena API and any framework (e.g. Java EE, GSW) for development client part of experimental prototype. The key semantic search component has been got.

***Keywords**: semantic search, social profile, social indexing of documents, social network.*