УДК 519.7:007.52

О.Ю. Тимофеева

Харьковский национальный университет радиоэлектроники, Харьков

РАЗРАБОТКА МЕТОДА РАСЧЕТА СОЦИАЛЬНОЙ ДИСТАНЦИИ В WEB 2.0

В статье рассмотрена задача разработки метода расчета социальной дистанции между пользователями Интернет-ресурсов в Web 2.0, использующего в качестве входных данных профили пользователей, построенные на основе онтологий. Данный подход позволяет учитывать семантику данных, описывающих пользователя для максимально точного сравнения. Предлагаемый метод может быть использован при разработке социальных сетей, блогов и систем электронной коммерции.

Ключевые слова: *социальная дистанция, модель представления пользователей, семантическое сравнение, онтология.*

Введение

На данный момент задача сравнения пользователей Интернет-ресурсов является особенно актуальной. Необходимость сравнивать пользователей между собой и находить группы похожих пользователей присутствует как в электронной коммерции, так и при разработке различных блогов, социальных сетей и сайтов знакомств.

В настоящее время данная задача решается в основном через представление пользователей посредством признакового описания и использование простой фильтрации по заданным признакам (например, фильтрация по возрасту). Такой метод можно использовать при работе с конечным числом заполняемых явно признаков (т.е. тех, которые пользователь указал сам). В то же время подобная фильтрация задает слишком жесткие ограничения, в результате которых могут отсекаться профили тех людей, которые могли бы быть интересными данному пользователю. Кроме того, подобным образом можно обрабатывать данные социально-демографического характера, но неудобно работать с информацией, касающейся привычек и предпочтений пользователя.

Главная проблема, встречающаяся на пути к попытке применения интеллектуальных систем рекомендаций в данной области – это отсутствие эффективных методов семантического сравнения пользователей.

Постановка задачи. В то время как в Web 2.0 предполагается широкое использование онтологий для представления знаний (в том числе и знаний о пользователях), имеет смысл разработать метод расчета социальной дистанции между объектами на основе онтологий.

Целью работы является разработка метода расчета социальной дистанции между различными пользователями одного ресурса на основе семантической близости их интересов и предпочтений. Исходными данными являются профайлы пользователей, построенные по модели с использованием онтологического представления концептов.

Расчет социальной дистанции

Социальная дистанция, устанавливающая степень близости и сходства людей между собой, зависит от признаков, их описывающих. Если два человека совпадают по большей части из этих признаков, можно сказать, что социальная дистанция между ними невелика.

Задача вычисления социальной дистанции сводится к задаче поиска расстояния между двумя точками в n-мерном пространстве, где в виде точек представляются профили пользователей, а n – это количество свойств, которыми они описываются. Это значит, что дистанция между двумя пользователями будет находиться с помощью сложной функции вида

$$D_{xy} = \sqrt{\sum_{k=1}^{n} d(p_{xk}, p_{yk})^2}, \qquad (1)$$

где D_{xy} – социальная дистанция между двумя пользователями; k – порядковый номер свойства в профиле; p_{xk}, p_{yk} – значение k-го свойства у двух пользователей; $d(p_{xk}, p_{yk})$ – компонентная формула для вычисления расстояния между первым и вторым объектом по k-му свойству.

В зависимости от типа данных k-го свойства, расстояние $d(p_{xk}, p_{yk})$ будет считаться по-разному. Для более корректной обработки данных также необходимо провести нормировку – корректировку значений в соответствии с некоторыми функциями преобразования с целью сделать их более удобными для сравнения. Если нормировку не проводить, это может быть чревато тем, что свойства, описывающиеся числовыми значениями более высокого порядка будут иметь больший вес, чем свойства, описывающиеся числами меньшего порядка. Поскольку в большинстве случаев это будет некорректно, вследствие чего это отрицательным образом отразиться на результате решения задачи, предлагается проводить нормировку

значений расстояний $d(p_{xk}, p_{yk})$ для всех свойств объектов так, чтобы нормированное значение находилось в интервале [0;1].

В этом случае расстояние $d(p_{xk}, p_{yk})$ будет принимать значение, равное нулю в том и только в том случае, если объекты p_{xk} и p_{yk} совпадают, то есть значения всех свойств, их описывающих, равны.

В случае работы с простыми числовыми типами данных (например, описывающих возраст или средний доход пользователя), где каждое свойство описывается только одним числом, $d(p_{xk}, p_{yk})$ будет вычисляться как модуль разности двух значений. Для осуществления нормировки в таком случае необходимо установить граничные возможные значения для данных свойств, и значение $d(p_{xk}, p_{yk})$ поделить на разность между двумя граничными значениями.

При работе же с категориальными переменными, которыми также могут описываться свойства, необходимо определиться, каким именно образом будет вычисляться расстояние между ними. Один из способов вычисления такого расстояния представлен формулой

$$d(p_{xk}, p_{yk}) = \begin{cases} 0 \mid p_{xk} = p_{yk} \\ 1 \mid p_{xk} \neq p_{yk} \end{cases}. \qquad (2)$$

Данной формулой удобно пользоваться тогда, когда один атрибут может принимать два возможных значения. Также эта формула применима при работе с данными логического типа.

Вычисление компонентной формулы $d(p_{xk}, p_{yk})$ наиболее усложнено при работе не с простыми свойствами, а со сложными – когда значением свойства одного объекта могут выступать объекты другого класса. Задача усложняется еще больше, когда допустимо несколько значений для одного свойства. Следовательно, необходимо разработать метод вычисления семантического расстояния между значениями определенных свойств объектов, который работал бы одинаково успешно вне зависимости от специфики предметной области, которую затрагивает данное свойство.

Если рассматривать представление свойств пользователей в виде онтологий, то можно выделить два основных подхода к вычислению семантической близости между любыми двумя ее концептами. Первый способ основывается на использовании концептуальной схемы предметной области и сводится к нахождению расстояния между двумя узлами графа. Применение такого подхода особенно удобно, когда граф не имеет циклов и между любой парой вершин его существует ровно один путь, то есть граф является деревом. Используя терминологию онтологического инжиниринга, это означает, что каждый класс и каждый экземпляр класса является наследником лишь одного суперкласса, что в большинстве случаев случается при построении онтологии на основе формальной таксономии.

В таком случае расстояние считается как количество шагов в кратчайшем пути между двумя узлами графа (если граф невзвешенный). Если граф является деревом, то между любыми двумя узлами существует только один путь, и нахождение его представляется достаточно простой задачей.

Второй способ основан на использовании свойств экземпляров классов при нахождении расстояния между ними. В этом случае сравнивается значение одного и того же свойства двух разных экземпляров; если значения совпадают, то расстояние между экземплярами по данному свойству будет равно нулю. В противном же случае, если значения не совпадают, оно будет равняться единице.

Исходя из поставленной задачи, можно совмещать оба подхода так, чтобы это в результате упрощало ее решение. При использовании онтологического подхода, информация, описывающая понятия, содержится как в их свойствах, так и в самой решетке понятий.

Кроме этого, важно то, что значения одного и того же признака, описывающего понятие, можно выражать и через свойства класса, и с помощью связей в онтологии.

Для более детального сравнения двух понятий (экземпляров или классов) необходимо использовать информацию, представленную в обеих формах. То есть, при вычислении расстояния между двумя экземплярами онтологии будет вычисляться как расстояние по графу, описывающему концептуальную схему онтологии, так и расстояние на основе попарного сравнения значения свойств экземпляров.

$$d(a,b) = \sqrt{\left(\frac{1}{2}\right)^2 + \sum_{k=1}^{n} d(v_{ak}, v_{bk})^2}, \qquad (3)$$

где $d(a,b)$ – расстояние между двумя экземплярами онтологии; l – количество шагов в кратчайшем пути между двумя узлами графа; v_{ak} – значение k-го свойства у первого экземпляра; v_{bk} – значение k-го свойства у второго экземпляра.

Для нормировки значения расстояния между сложными свойствами, необходимо для каждого свойства найти такую пару значений, расстояние $d(a,b)$ между которой было бы максимальным.

Исходя из онтологического представления значений свойств, можно сказать, что у данных двух экземпляров все значения свойств, их описывающих, будут разными, и расстояние по графу, соответствующему решетке понятий онтологии, будет максимальным. В этом случае нормированное значения расстояния между двумя экземплярами будет равно

$$d(a,b)=\frac{\sqrt{\left(\frac{1}{2}\right)^2+\sum_{k=1}^{n} d(v_{ak},v_{bk})^2}}{d_{max}(a,b)}. \qquad (4)$$

Стоит отметить, что при изменении структуры онтологии (при ее расширении и добавлении новых классов или свойств), расстояние между максимально удаленными экземплярами необходимо будет пересчитывать.

Выше был рассмотрен способ вычисления семантического расстояния между двумя понятиями, представленными как экземпляры классов онтологии. Учитывая, что ряд свойств класса "пользователь" может иметь не одно, а несколько значений, необходимо на основании формулы (4) разработать метод вычисления расстояния между такими свойствами.

Поскольку предвидеть заранее количество значений того или иного свойства у каждого из пользователей невозможно, а ограничивать пользователя строго заданным количеством значений нежелательно, формула вычисления семантического расстояния между значениями свойств должна содержать число значений данного свойства для каждого из сравниваемых пользователей.

Модифицированная формула вычисления семантического расстояния между всеми значениями данного свойства для двух пользователей вычисляется по формуле

$$d(p_{xk},p_{yk})=\frac{\sqrt{\left(\frac{1}{2}\right)^2+\sum_{k=1}^{n} d(v_{ak},v_{bk})^2}}{d_{max}(a,b)*n_1*n_2}, \qquad (5)$$

где n_1 – количество значений данного свойства у первого пользователя; n_2 – количество значений свойства у второго пользователя.

В этом случае число n_1*n_2 – это количество связей между всеми значениями данного свойства, взятыми попарно. Таким образом, добавив это число в знаменатель, мы получим среднее значение расстояния между каждой парой значений свойства для данных двух пользователей, что позволит сделать вывод о том, насколько разнятся интересы этих пользователей в данной сфере.

Выводы

Предложен метод расчета социальной дистанции между пользователями интернет-ресурсов на основе семантического сравнения их социально-демаграфических данных, а также их интересов и предпочтений. Данный метод использует профайлы пользователей, построенные на основе онтологий, что позволяет осуществить семантический подход к решению поставленной задачи.

Список литературы

1. Governor J., Nickull D. Web 2.0 Architectures. – Sebastopol: O'Reilly, 2009. – 272 p.

2. Segaran T. Programming Collective Intelligence. – Sebastopol: O'Reilly, 2007. – 368 p.

3. Zhang D., Lee W. S. Web taxonomy integration using support vector machines. – New York:ACM, 2004. – 481 p.

4. Giunchiglia F. Semantic schema matching / F. Giunchiglia, M. Yatskevich, P. Shvaiko. – Heidelberg: Springer, 2005. – 635 p.

5. Bell G. Building Social Web Applications. – Sebastopol: O'Reilly, 2009. – 448 p.

6. Boyd D., Ellison N. Social Network Sites: Definition, History, and Scholarship. – Journal of Computer-Mediated Communication, Vol. 13, №1: 2007. – P. 210-230.

7. Doan A., Madhavan J., Domingos P., Halevy A. Learning to map between ontologies on the semantic Web. – New York: ACM, 2002. – 662 p.

8. Canny J. Collaborative Filtering With Privacy via Factor Analysis.– Proceedings of the 2002 IEEE Symposium on Security and Privacy: Tampere, Finland, 2002. – P. 238-245.

9. Mitchell T.M., Carbonell J. G., Michalski R. S. Machine Learning: A Guide to Current Research. – New York: ACM, 1998. – 432 p.

10. Giunchiglia F., Shvaiko P. Semantic matching. – New York: Cambridge University Press, 2003. – 280 p.

Поступила в редколлегию 1.03.2011

Рецензент: д-р техн. наук, проф. В.Я. Терзиян, Харьковский национальный университет радиоэлектроники, Харьков.

РОЗРОБКА МЕТОДУ РОЗРАХУНКУ СОЦІАЛЬНОЇ ДИСТАНЦІЇ У WEB 2.0

О.Ю. Тимофієва

У статті розглянуто завдання розробки методу розрахунку соціальної дистанції між користувачами Інтернет-ресурсів в Web 2.0, що використовує як вхідні дані профілі користувачів, побудовані на основі онтологій. Даний підхід дозволяє враховувати семантику даних, що описують користувача для максимально точного порівняння. Пропонований метод може бути використаний при розробці соціальних мереж, блогів і систем електронної комерції.

Ключові слова: *соціальна дистанція, модель представлення користувачів, семантичне порівняння, онтологія.*

METHOD FOR CALCULATION SOCIAL DISTANCE FUNCTION FOR WEB 2.0 DEVELOPMENT

O.Yu. Tymofiieva

In the article the task of development the method of calculation of social distance is considered between the users of Internetresources in Web 2.0, that uses as entrance these types of users, built on the basis of ontologys. This approach allows to take into account semantics of information which describe an user for maximally exact comparison. The offered method can be used for development of social networks, blogs and systems of electronic commerce.

Keywords: *social distance, model of presentation of users, semantic comparison, ontology.*