ФОЛКСОНОМИЯ КАК ИСТОЧНИК ДЛЯ ПОСТРОЕНИЯ ОНТОЛОГИЙ И АССОЦИАТИВНЫХ СЛОВАРЕЙ. ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ

Канищева О.В.
Национальный технический университет "Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/2, тел. 707–63–60,
e-mail: olya-kanisheva@rambler.ru

С развитием концепции семантического веба – web 2.0 появились и активно развиваются такие социальные ресурсы как Flickr [1], Instagram [2], Picasaweb [3], Photobucket [4] и другие. В основе таких систем лежит методика проектирования, в результате которой с помощью сетевых взаимодействий, системы становятся тем лучше, чем больше людей ими пользуются. Таким образом пользователи (не профессионалы) добавляют свои собственные ключевые слова (признаки) к информационным объектам (фото, видео и др.). Информации (тегов) становится все больше, однако вопросы о её надёжности, достоверности и объективности подвергаются сомнениям.

В результате создания таких ресурсов и их разметки происходит так называемая *фолксономия* (англ. folksonomy, от folk – народный + taxonomy таксономия, от гр. расположение по порядку + закон) – народная классификация, практика совместной категоризации информации (текстов, ссылок, фото, видео клипов и т. п.) посредством произвольно выбираемых меток, называемых тегами [5].

Системы, реализующие маркировку, состоят из трех главных элементов: пользователи, ресурсы и признаки (теги). Хотя в большинстве систем теги не являются обязательными, но они являются очень важными, благодаря этим меткам у пользователей есть возможность осуществлять поиск, сортировать и объединять в группы сходные объекты.

Однако никакими правилами или словарями при разметке ресурсов пользователи не пользуются. Это влечет за собой орфографические ошибки, мультиязычное написание слов (*muenchen, munich*), использование различных символов при написании (*ny, new_york, “new york”, newyork*), синонимы, флективные формы слов (множественное число и др.), эмоциональную лексику (*pretty, nice*), местоимения и др. Весь этот шум мешает качественно проводить кластеризацию информационных объектов и обработку ключевых слов.

Примеры тегов для произвольного изображения *1) niagara, falls, buffalo, new, york, ny, city, nature, water, steam, haze, park, state, erosion, historic, tourist, boat; 2) panda, giant, wildlife, nature, bear, panda bear, national zoo, The Perfect Photographer, AnimalKindomElite, SpecAnimal*.

Особенностью фолксономии является то, что правил для маркировки ресурсов для пользователей как таковых нет. Пользователи могут как присвоить ресурсу одно ключевое слово, так и 10 слов. Если ресурсы похожи и относятся к одному пользователю, то он копирует теги и меняет несколько слов, а в неко-

торых случаях так их и оставляет. В работе [2] авторы показали, что при анализе сервиса Delicious за 2004 год самыми популярными тегами были: *software, design, programming, music, politics, web, news, blog, css, linux, art, osx, java, mac, blogs, reference, fun, python, games, tech, photography, humor, tools, delicious, rss, firefox, toread, comics*. Многие из них являются техническими тегами, которые отражают технологические интересы (*rss, firefox, python, java, linux*). Некоторые из них описываю жанр или носят описательный характер (*comics, humor, fun, photography*). Одним из интересных тегов является *toread,* он используется для самоорганизации и напоминания. Тег wishlist (http://del.icio.us/tag/wishlist) использовался пользователями для того, чтобы выдвинуть на первый план потребительские группы, которыми они были заинтересованы [6].

Существует достаточно много работ, связанных с фолксономией, часть из них посвящено классификации в фолксономии, созданию онтологий, анализу лексики, которую пользователи используют при аннотировании ресурсов и т.д.

В связи со спецификой тегов появляются программные разработки, которые позволяют произвести автоматическую разметку изображения. Однако и они не лишены недостатков, так как напрямую связаны с задачей распознавания образов.

Основные подходы, которые используются для обработки информации в фолксономиях – это статистические методы, методы вычисления семантической близости (расстояние Левенштейна, коэффициент Жаккара и др.) [7, 8]. В большинстве подходов используются различные лингвистические ресурсы: WordNet, Wikipedia, DBpedia, YAGO, ConceptNet и др. Однако эти ресурсы не совершенны для решения проблем обработки ключевых слов, так как не охватывают весь спектр лексики, представленной в фолксономиях.

Особое место занимает очистка тегов и приведение их к унифицированному виду. Для этого используют расстояния Левенштейна, стемминг, лемматизацию и словари-тезаурусы. После этапа очистки можно находить семантически близкие теги и проводить кластеризацию ресурсов.

Список литературы

1. https://www.flickr.com/
2. https://instagram.com/
3. https://picasaweb.com/
4. http://photobucket.com/
5. http://en.wikipedia.org/wiki/Folksonomy
6. Adam Mathes (2004) *Folksonomies – Cooperative Classification and Communication Through Shared Metadata*. Access: http://www.adammathes.com/academic /computer-mediated-communication/ folksonomies.html
7. Lucia Specia, Enrico Motta (2007) *Integrating Folksonomies with the Semantic Web*. In: The Semantic Web: Research and Applications Lecture Notes in Computer Science Volume 4519, 2007, pp. 624-639.
8. Chao Wu, Bo Zhou (2009) *Semantic Relatedness in Folksonomy*. In: International Conference on New Trends in Information and Service Science, pp. 760 – 765.