Сравнение онлайн-сообществ на основе лексического анализа ленты новостей

Д. А. Усталов[1], Ф. В. Краснов[2], Р. Э. Яворский[3]

[1] dau@imm.uran.ru, [2] fk@sk.ru, [3] ryavorsky@sk.ru

[1] ИММ УрО РАН, Екатеринбург, Россия
[2,3] Фонд Сколково, Москва, Россия

Аннотация. Предложен подход к формальной оценке характеристик онлайн-сообществ путём лексического анализа содержания ленты новостей. Описаны результаты экспериментов по обработке содержания новостных лент трёх разных онлайн-сообществ: IT-EBURG, GIS-Lab, Сколково. Такие параметры, как количество терминов и доля профессиональной лексики в новостях, рассматривается как важная характеристика культурной среды исследуемого онлайн-сообщества.

Ключевые слова: извлечение ключевых слов; профессиональное сообщество; онлайн-сообщество; характеристики онлайн-сообществ.

Введение

Онлайн-сообщество — это группа людей, объединённых на базе Web-сервиса, предоставляющего функциональность онлайн-дискуссии и совместной работы.

Несмотря на внешнюю схожесть, онлайн-сообщества заметно отличаются друг от друга по целям, внутренней структуре, принятому стилю общения, и другим характеристикам.

Разработка методов и инструментов для определения основных характеристик онлайн-сообщества является интересной и востребованной задачей [1, 2, 3].

1 Цели и задачи работы

Цель работы состоит в том, чтобы найти формальные (с прозрачным алгоритмом расчёта) параметры онлайн-сообществ, характеризующие стиль общения, принятый среди его участников.

В рамках данного исследования взяты тридцать последовательных текстов за осень 2012 г. из ленты новостей трёх разных онлайн-сообществ:

— IT-EBURG — локальное сообщество IT-специалистов Екатеринбурга [4];
— GIS-Lab — профессиональное сообщество исследователей и разработчиков геоинформационных систем [5];
— онлайн-сообщество инновационного центра Сколково [6].

Выбор данного временного периода обусловлен необходимостью минимизировать зашумлённость данных новогодними праздниками и летними отпусками.

2 Реализация

Из каждой новости автоматически извлечены ключевые слова и словосочетания при помощи инструментария, описанного в [7].

Полученные ключевые слова и словосочетания были размечены экспертом на два класса: 1) общая лексика; 2) профессиональная лексика.

Для каждого текста посчитано количество ключевых слов и вычислена доля профессиональной лексики.

Стоит отметить, что в каждом сообществе, независимо от профессиональной принадлежности, в новостях встречаются объявления общего типа: информация о мероприятиях, анонсы, отчёты, и т. д.

Для компенсации этой особенности из выборки были исключены тексты, в которых доля профессиональной лексики не превышает 10%. Таких текстов оказалось всего 13 для IT-EBURG, 19 для GIS-Lab, 14 для Сколково.

3 Результаты и их анализ

Итоговые результаты отражены на схеме (рис. 1), где каждому маркеру соответствует одна запись в новостной ленте. Также отобра-

жены маркеры для усредненного значения параметров по каждому сообществу.

На схеме видно, что сложившиеся специализированные сообщества (IT-EBURG и GIS-Lab) отличаются большой долей профессиональных терминов и небольшим количеством ключевых слов, входящих в текст новостей. В сообществе Сколково явным образом выделяется большое количество терминов общей лексики.

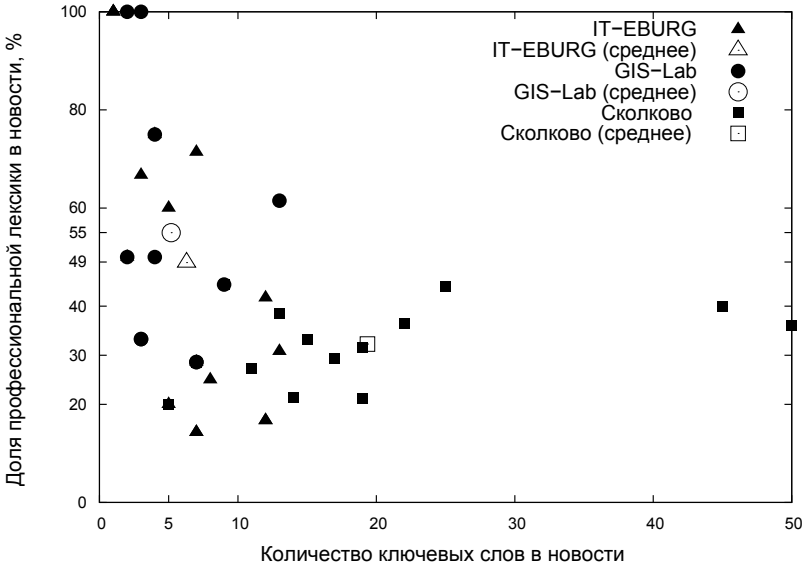

Рис. 1. Результаты экспертной оценки текстов новостей трёх онлайн-сообществ

Заключение

Предложен подход формальной оценки онлайн-сообщества путём лексического анализа новостных лент.

В этом направлении сделаны первые шаги. Отчасти подтвердилась гипотеза о том, что в качестве одного из таких параметров можно рассматривать долю профессиональной лексики в ключевых словах новостной ленты сообщества.

Для устоявшихся профессиональных сообществ этот параметр близок к 50% (49% для IT-EBURG и 55% для GIS-Lab), для проходящего стадию формирования онлайн-сообщества Сколково это параметр заметно ниже и составляет в среднем 30%.

Главным недостатком реализации предлагаемого подхода является наличие этапа ручной разметки, который ограничивает возможности масштабирования подхода.

В качестве следующего шага целесообразно использовать доступные средства автоматической тематической классификации документов [8].

Список источников

1. Yavorskiy R. Research Challenges of Dynamic Socio-Semantic Networks // The International Workshop on Concept Discovery in Unstructured Data (CDUD 2011). pp. 119–121.
2. Краснов, Ф. Развитие через общение, Intelligent Enterprise, т. 9, С. 18–21, 2012.
3. Краснов, Ф., Яворский, Р. Измерение уровня зрелости профессионального сообщества // Принята к публикации в журнале «Бизнес-информатика», 2013.
4. Новости — Информационные технологии в Екатеринбурге [Электронный ресурс]: Информационные технологии в Екатеринбурге. — Режим доступа: http://it-eburg.com/ (дата обращения: 05.02.2013)
5. GIS-Lab: Геоинформационные системы и Дистанционное зондирование Земли [Электронный ресурс]: GIS-Lab. — Режим доступа: http://gis-lab.info/ (дата обращения: 05.02.2013)
6. Infocenter [Электронный ресурс]: Инновационный центр Сколково. — Режим доступа: http://community.sk.ru/press/ (дата обращения: 05.02.2013)
7. Усталов, Д. Извлечение терминов из русскоязычных текстов при помощи графовых моделей / в сб. «Теория графов и приложения = Graphs theory and applications: материалы конференции». С. 62–69. — Екатеринбург: Изд-во Урал. ун-та, 2012.
8. Агеев, М. и др. Экспериментальные алгоритмы поиска/классификации и сравнение с «basic line» // Российский семинар по Оценке Методов Информационного Поиска (РОМИП 2004) — Пущино. – 2004. – С. 62–89.