

**ПОЛУАВТОМАТИЧЕСКАЯ ПРИВЯЗКА
СОВМЕСТИМЫХ АКСЕССУАРОВ К ТОВАРАМ
ИНТЕРНЕТ-МАГАЗИНА**

Зимовнов Андрей Вадимович

Аспирант

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: zimovnov@gmail.com

Для интернет-магазина, продающего товары из разделов электроники и бытовой техники, где наценка минимальна, дополнительную прибыль приносят продажи аксессуаров (например, чехлов для телефонов). Аксессуары отличаются высоким уровнем наценки, а следовательно стимулирование совместных продаж устройств с аксессуарами может принести ощутимую прибыль.

Эксперименты с автоматическим подбором аксессуаров показали, что качество предложений недостаточно для более агрессивного продвижения аксессуаров. Ручная привязка тысяч устройств к тысячам аксессуаров выглядит как перебор миллионов пар. Для ускорения процесса ручной привязки был разработан алгоритм, который для каждого аксессуара выводил 10 самых подходящих устройств. Ассессору предлагалось из предложенных 10 устройств выбрать те, с которыми действительно совместим рассматриваемый аксессуар. Это свело задачу к рассмотрению человеком десятков тысяч пар, вместо миллионов.

Для каждого устройства и аксессуара имелось текстовое описание, состоящее из разных полей (название, производитель, аннотация, и другие). В тексте описания аксессуаров в том или ином виде упоминались названия устройств. Однако задача усложнялась использованием *перечислений* (например, строчка «для iPad 1/2 / iPhone 3Gs/4/4S» понятна человеку, но автоматический алгоритм может ошибиться с привязкой устройства «iPad 4»), использованием *кратких названий* устройств (например, «Note 8» вместо «Samsung Galaxy Note 8.0»), использованием *синонимов* (например, «Galaxy S3» вместо «Samsung Galaxy S III»).

Проблемы *перечислений* и *кратких названий* исключили использование лишь одного простого признака «мощность пересечения» описаний устройства и аксессуара. Для достижения лучшего качества были также реализованы известные поисковые признаки f_i , которые учитывают популярность слов, порядок и расстояние между общими словами описаний [1, 4]. Проблема *синонимов* была решена

написанием нескольких регулярных выражений, которые покрыли большинство случаев, сложные случаи для популярных устройств (например, использование «iPad 5» вместо «iPad Air») были внесены как ручные правила замены.

В качестве функционала качества рассматриваемой задачи ранжирования был выбран MAP@10 [1]. Линейная комбинация поисковых признаков $r = \sum_{i=1}^k w_i f_i$ настраивалась методом SVM [2] на обучающую выборку. Обучающая выборка была получена разметкой 10% ассортимента аксессуаров, для каждого аксессуара предлагалось разметить 100 устройств с максимальным значением r при весах $\vec{w} = \vec{1}$. Итоговый алгоритм имел качество 0,92 на всей выборке, то есть настроенная линейная комбинация на практике часто показывала совместимые устройства в первых 10. Для ассессорской разметки несколькими людьми был реализован Web-интерфейс [3].

Работа выполнена при финансовой поддержке РФФИ, проект 14-07-00965.

Литература

1. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze. Introduction to Information Retrieval. 2008. Cambridge University Press, New York, NY, USA.
2. Yisong Yue, Thomas Finley, Filip Radlinski, and Thorsten Joachims. A support vector method for optimizing average precision // In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '07). 2007. ACM, New York, NY, USA, 271-278.
3. Документация Web-фреймворка Flask для Python:
<http://flask.pocoo.org/docs>
4. Документация поискового движка Sphinx:
<http://sphinxsearch.com/docs>