FashionAsk: Pushing Community Answers to Your Fingertips

Wei Zhang, Lei Pang, Chong-Wah Ngo
Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
{wzhang34, leipang3}@student.cityu.edu.hk, cscwngo@cityu.edu.hk

ABSTRACT

We demonstrate a multimedia-based question-answering system, named *FashionAsk*, by allowing users to ask questions referring to pictures snapped by mobile devices. Specifically, instead of asking verbose questions to depict visual instances, direct pictures are provided as part of questions. To answer these multi-modal questions, FashionAsk performs a large-scale instance search to infer the names of instances, and then matches with similar questions from community-contributed QA websites as answers. The demonstration is conducted on a million-scale dataset of Web images and QA pairs in the domain of fashion products. Asking a multimedia question through FashionAsk can take as short as five seconds to retrieve the candidate answer as well as suggested questions.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

General Terms

Algorithms, Performance, Experimentation

Keywords

Multimedia question answering, instance naming, question matching

1. INTRODUCTION

Community-contributed question-answering (cQA) websites, such as Yahoo! Answers[1], has gained popularity in the past few years, and accumulated hundred of millions question-answer (QA) pairs. Most of the QA pairs are text-based, where a user posts question by texting and expects text answers in return by search or by other users. With the convenient use of mobile devices for snapping pictures, there is a shift recently on how community users post questions. For instance, instead of asking a question by providing a long text description for a visual instance, direct visual examples are included as part of the question. For example, ask the question "Who is the designer of this rock-and-roll t-shirt?" by providing a picture of the t-shirt. In this demo, a system, named FashionAsk, capable of answering multimedia-based questions in the domain of fashion is demonstrated. Specifically, a question is answered by first naming the visual instance in the given picture, e.g., "Twisted Sister Rock-and-Roll t-shirt", through a large-scale search of Web images. The original text question is then augmented with the extracted name entity, and searched against cQA archives. To this end, the system returns potential answers in text, as well as shortlisting a few suggested QA pairs. Our current system operates on a dataset with 1.1 millions Web images and 1.5 millions QA pairs.

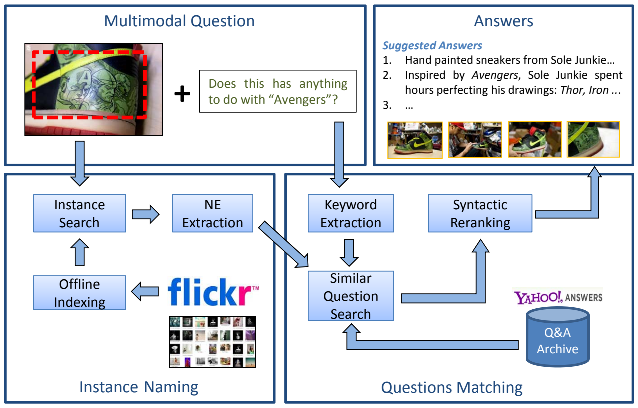


Figure 1: Framework of FashionAsk.

Compared to existing cQA websites, which rely on text for searching similar QA pairs, our system is novel by enabling the answering of multimodal questions with visual content analysis. Compared to recent mobile search engines, such as Google Goggles[2] and SnapTell[3], the developed system is capable of handling instances of non-rigid shapes and non-planar surfaces, such as t-shirts, sneakers, and handbags. Furthermore, instead of returning webpages to users, our system is more specific by allowing users to expect answers and suggested questions that might directly address the information needed.

2. TECHNOLOGY

Figure 1 shows the framework of FashionAsk. The system is composed of two components: *instance naming* and *QA*

[1] http://answers.yahoo.com/

[2] http://www.google.com/mobile/goggles/#text

[3] http://www.snaptell.com/

MM'12, October 29–November 2, 2012, Nara, Japan.
ACM 978-1-4503-1089-5/12/10.

matching. The former infers the name of a fashion instance, while the latter searches for similar QA pairs.

2.1 Instance Search and Naming

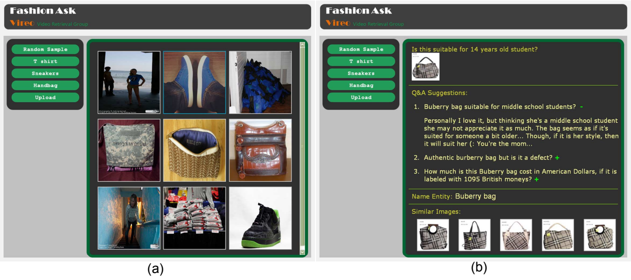
The basic idea is to search similar images of the given instance, and then infer the name of the instance directly from metadata of these images. The main challenges are that fashion instances could appear in different forms and exist in different backgrounds, as shown in Figure 2(a). Thus, our framework considers a spatial verification of instance by triangulation matching to address the challenges, which will be briefly elaborated next.

Instance search is based on Bag-of-visual-Words model [2]. First, a hierarchical vocabulary tree with one million leaf nodes is built offline. To enable robust search, each word is indexed with its spatial location and short Hamming signature [1]. To support efficient large-scale search, the index files are distributed to several servers. During online search, a query is matched against images in the dataset by traversing the inverted file with Hamming signature verification. For each matched image, triangulation meshes are further constructed to characterize the spatial consistency. Our technique is different from other techniques such as WGC [1] and RANSAC, which assume an explicit transformation model. In contrast, the triangulation-based matching does not assume such models, but locally measures the relative positioning of visual words to derive coherency scores for image ranking. The matching is found to be more effective than other methods in handling objects with non-rigid shapes and non-planar surfaces.

With the list of retrieved images, name entity extraction (NE) is processed by parsing the metadata (title and description) of each image with the Berkeley Parser[4]. The noun phrases are then extracted directly as candidate names. The likelihood of a phrase being the name of the instance is measured by its phrase frequency weighted by the coherency scores of instance search.

2.2 Questions Matching

As questions in cQA websites are asked in natural language, similar questions are varied in terms of lexical, syntactic, and semantic features. Hence, retrieving similar questions is not trivial, and requires natural language processing. For the consideration of speed, a two-stage question matching method is developed. In the first stage, the keywords extracted from user's question together with top name candidates are used to retrieve a small set of similar questions from a QA dataset with the BM25 retrieval model, which ranks documents based on a probabilistic model. In the second stage, the original question is augmented by replacing the pronoun, such as "this", with candidate names, and matched against the set of retrieved questions. The matching technique is implemented based on the syntactic tree matching algorithm [3], which further considers grammatical structures of questions. The end result is a ranked list of QA pairs that are presented as the answer and suggested questions for users.



Is this suitable for 14 years old student?

Q&A Suggestions:

1. Buberry bag suitable for middle school students? -

   Personally I love it, but thinking she's a middle school student she may not appreciate it as much. The bag seems as if it's suited for someone a bit older... Though, if it is her style, then it will suit her (: You're the mom...
2. Authentic burberry bag but is it a defect? +
3. How much is this Buberry bag cost in American Dollars, if it is labeled with 1095 British moneys? +

Name Entity: Buberry bag

Similar Images:

Figure 2: Interface for FashionAsk: (a) showing examples of fashion pictures in our dataset; (b) asking a question by picking or uploading a picture, and followed by the answer and suggested questions returned by our system.

3. SYSTEM AND USER INTERFACE

Interface. The system can be accessed through a Web browser operated either on a PC client or a mobile phone. As the interface shown in Figure 2(a), a user can either upload a picture, or pick an image randomly listed by our system, and then ask a fashion related question. Users can also crop the region-of-interest from an image when issuing questions. The returned page will show the name of the visual instance, the most likely answer, and a few suggested questions retrieved from our database, as illustrated in Figure 2(b).

Performance. The system currently consists of 1.1 millions Web images, crawled from Flickr by querying several fashion related keywords (e.g., t-shirt, sneaker, and handbags) and 142 popular tags. Our dataset also has 1.5 millions QA pairs crawled from Yahoo! Answers, one quarter of which are related to fashion, and the others are randomly crawled as distractors. Based on current setup, the performance of our system is as follows when testing on 180 multimedia questions involving 50 fashion instances: the mean average precision of instance search is 18.8%; the accuracy of finding the right instance name as the top-1 candidate is 21.7%; and the accuracy of returning a right answer or a related question in top-10 list is 19.8%. The system is currently run on two machines with 2.67GHz CPU and 16GB main memory. By current setup, instance naming consumes 1.7 seconds, and question matching takes about 2.9 seconds to rank similar questions. In total, the user needs to wait approximate five seconds to get the result. The time could be within one second, if more than ten machines are employed to parallelize instances searching and questions matching.

4. ACKNOWLEDGMENTS

The work described in this paper was fully supported by two grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 119610 and CityU 118812).

5. REFERENCES

[1] H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, 2008.

[2] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, volume 2, pages 1470–1477, Oct. 2003.

[3] K. Wang, Z. Ming, and T.-S. Chua. A syntactic tree matching approach to finding similar questions in community-based qa services. In *SIGIR*, 2009.

[4] http://code.google.com/p/berkeleyparser/