EntityTagger: Automatically Tagging Entities With Descriptive Phrases

Kaushik Chakrabarti, Surajit Chaudhuri, Tao Cheng, Dong Xin
Microsoft Research
One Microsoft Way
Redmond, WA 98052
{kaushik, surajitc, taocheng, dongxin}@microsoft.com

ABSTRACT

We consider the problem of *entity tagging*: given one or more named entities from a specific domain, the goal is to automatically associate descriptive phrases, referred to as *etags* (entity tags), to each entity. Consider a product catalog containing product names and possibly short descriptions. For a product in the catalog, say *Ricoh G600 Digital Camera*, we want to associate etags such as "water resistant", "rugged" and "outdoor" to it, even though its name or description does not mention those phrases. Entity tagging can enable more effective search over entities. We propose to leverage signals in web documents to perform such tagging. We develop techniques to perform such tagging in a domain independent manner while ensuring high precision and high recall.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Design, Experimentation

Keywords

Entity tagging, tag discovery, tag association

1. INTRODUCTION

Consider an entity search engine like a product search engine. Often, users search for products based on desired features. Examples in the camera domain are [underwater disposable camera], [rugged camera] and [point and shoot camera]. We refer to such queries as *"search entity by feature" (SEF)* queries. A recent study reports that about 42% of all product queries are *SEF* queries [2].[1]

For *SEF* queries, searches over the product catalog often miss relevant results as the features that are searched on may not be included in the information about relevant products in the catalog [1]. Consider a product catalog containing the name, technical specifications and possibly a short description of each product. For query [rugged camera], *Ricoh G600 Digital Camera* is a good hit but its name, technical specifications or short description in the catalog might not have the mention of "rugged". Hence, search over the above catalog would fail to return this relevant product. If we can automatically assign descriptive phrases with high precision to entities (such as assigning "rugged", "outdoor" and "water resistant" to *Ricoh G600 Digital Camera*) and augment the product catalog with them, search over the catalog will be able to answer such *SEF* queries more effectively.

Tagging has become very popular in Web 2.0 sites (e.g., Flickr) where users *manually* annotate items like images, videos and internet bookmarks with phrases to enable effective browsing and search. However, unlike tags in Web 2.0 systems that are free-form and have no fixed semantics, for entities, we focus on a restricted class of tags that can be *automatically* identified: *those that describe features of the entity* (e.g., "rugged", "outdoor" and "water resistant" for the entity *Ricoh G600 Digital Camera*). We refer to such tags as *entity tags* or *etags* in short. We focus on this restricted class for three reasons. First, we believe such tags would be most helpful in answering *SEF* queries. Second, we show it is feasible to *automatically* identify such tags in a domain-independent manner with high quality instead of the manual tagging as practiced in Web 2.0 sites. Third, we can systematically evaluate the quality of our entity tagging system for the above class of tags.

To perform entity tagging, we leverage many resources including reviews, blogs and expert advice sites that contain rich information about entities of interest. Our approach is to derive sufficiently robust signals from these resources to identify such entity tags automatically. The technical challenge is to identify such entity tags in a *domain-independent manner* while ensuring *high precision and high recall*.

2. ARCHITECTURE AND TECHNIQUES

We develop a novel, two-step architecture that uses a judicious combination of precise lexical patterns and large scale, co-occurrence analysis over web documents to associate etags to entities. The architecture is shown in Figure 1. We describe the two steps in further detail:

Step 1: Etag Discovery: In the first step, we use robust and precise lexical patterns that identify etags for an entity *domain* (e.g., for the camera domain). Note that we discover these tags without reference to any specific entity. For example, for the camera domain, we will discover tags like "rugged", "ultracompact" and "prosumer". Our system only requires domain experts to provide a set of alternative strings used to denote a domain (i.e., domain names). For

[1] [2] reports that 19.9% of all web search queries are product queries out of which 8.56% are *SEF* queries (referred to as general product queries) while 11.35% are specific product name queries.

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

Domain	Entity	Associated Etags
Camera	Ricoh G600 Digital Camera	waterproof, outdoor, rugged, 10mp, ...
Camera	Go Photo Easy Pix 30 Digital Camera	pink, blue, ultra compact, mini, ...
Camera	Sony Cyber-shot DSC-W220 Point and Shoot Camera	image stabilized, digital zoom, compact, ...
Shirt	Ralph Lauren Childrenswear Striped Oxford Shirt	polo, soft, ...
Shirt	Dogwood Boys Striped Short Sleeve Polo	light blue, little boys, cotton, ...
Shirt	Under Armour Heatgear Short Sleeve Tee Girls	base layer, moisture wicking, casual, lightweight, ...

Table 1: Example Etags Associated to Entities

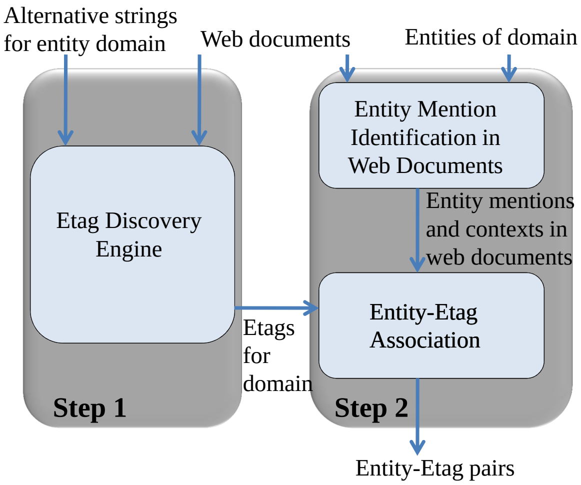
Alternative strings for entity domain
Web documents
Entities of domain
Etag Discovery Engine
Entity Mention Identification in Web Documents
Entity mentions and contexts in web documents
Entity-Etag Association
Etags for domain
Step 1
Step 2
Entity-Etag pairs

Figure 1: System Architecture

example, the alternative strings for the camera domain are {"camera", "digital camera"}. We look for patterns like "is a t $n_{\mathcal{D}}$", "t $n_{\mathcal{D}}$ such as" and "and other t $n_{\mathcal{D}}$" in web documents where t is a phrase, $n_{\mathcal{D}}$ is one of the alternative strings for the domain $\mathcal{D}$ and "t $n_{\mathcal{D}}$" is a noun phrase. For example, for the camera domain, we look for patterns like "is a *t* camera", "*t* cameras such as" and "and other *t* cameras". These patterns are inspired by Hearst patterns but differ from traditional Hearst patterns. If there are sufficient number of web documents containing the pattern, we identify t as an etag of the domain $\mathcal{D}$. We leverage the MapReduce/Dryad framework to discover etags scalably from web documents.

Step 2: Entity-Etag Association: In the second step, we associate the etags discovered for a domain with entities belonging to the domain. We leverage textual proximity between entities and etags in web documents to perform this association. We aggregate evidence across all web documents to achieve high precision and high recall. For example, if "rugged" is an etag discovered for the camera domain and it is mentioned near *Ricoh G600 Digital Camera* in many web documents, we will associate "rugged" to *Ricoh G600 Digital Camera*. This step has two software components:

(a) **Entity Mention Identification in Web Documents**: Since our association is based on text proximity between entities and etags, we first need to identify where the entities are mentioned in web documents. We refer to them as *entity mentions*. This component outputs not only the entity mentions but also their contexts. We develop techniques that are significantly more robust to approximate mentions than prior solutions. We develop algorithms that leverage the MapReduce/Dryad framework to scalably identify mentions of entities in web documents.

(b) **Entity-Etag Association Using Contexts**: Given the etags for a domain (discovered in Step 1) and the mentions of entities and their contexts in web documents (identified in Step 2(a)), we can look into the contexts to find out which entity-tag pairs appear in close proximity of each other. We aggregate evidence across all web documents using robust statistical techniques (based on G-test for goodness-of-fit) to associate etags to entities. This computation can be done scalably, by piggy-backing it with entity mention identification (Step 2(a)).

3. EXPERIMENTAL EVALUATION

Setup: Our experimental study uses a snapshot of the web corpus of high static rank documents, consisting of roughly 1.4 billion documents with total corpus size at 35.2T. We conduct thorough empirical evaluation on two real entity domains with very different characteristics, namely camera domain (with 3,557 camera entities) and shirt domain (with 22,182 shirt entities). The input alternative strings used for the camera domain are: "camera", "digital camera", and for the shirt domain are: "shirt", "t shirt".

Evaluation Method: We use the traditional notion of precision and recall for evaluating the entity tagging results. We perform expert judgement to evaluate the accuracy of the results produced by our system. We use the average number of etags associated to an entity as "recall", since we do not know the absolute set of etags for an entity.

Results: A few example entities with etags associated to them by our system are listed in Table 1. In Table 2 we report the number of etags discovered for these two domains respectively. The set of etags discovered are highly accurate, at around 99% precision.

Domain	Total Number of Etags Discovered
Camera	1,166
Shirt	935

Table 2: Statistics of Etags Discovered

Our empirical study show that our solution on average assigns ~10 etags per entity with ~85% precision and ~8 etags per entity with ~80% precision for the camera domain and shirt domain respectively. Compared with prior solutions, our system generates two orders of magnitude higher number of etags per entity on average while maintaining high accuracy. Furthermore, our algorithms scale well: we are able to perform both etag discovery and entity-etag association on all high static rank web documents on MapReduce/Dryad clusters in 3-4 hours.

4. REFERENCES

[1] S. Agrawal et al. Exploiting web search engines to search structured databases. In *WWW*, 2009.

[2] R. Kumar and A. Tomkins. A characterization of online search behavior. *Data Engineering Bulletin*, 2009.