

# A Web of Concepts

Nilesh Dalvi   Ravi Kumar   Bo Pang   Raghu Ramakrishnan   Andrew Tomkins  
Philip Bohannon   Sathiya Keerthi   Srujana Merugu  
Yahoo! Research  
701 First Ave.  
Sunnyvale, CA 94089, USA.

{ndalvi,ravikumar,bopang,ramakris,atomkins,plb,selvarak,srujana}

We make the case for developing a *web of concepts* by starting with the current view of web (comprised of hyperlinked pages, or documents, each seen as a bag of words), extracting concept-centric metadata, and stitching it together to create a semantically rich aggregate view of all the information available on the web for each concept instance. The goal of building and maintaining such a web of concepts presents many challenges, but also offers the promise of enabling many powerful applications, including novel search and information discovery paradigms. We present the goal, motivate it with example usage scenarios and some analysis of Yahoo! logs, and discuss the challenges in building and leveraging such a web of concepts. We place this ambitious research agenda in the context of the state of the art in the literature, and describe various ongoing efforts at Yahoo! Research that are related.

## Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Miscellaneous*

## General Terms

Experimentation, Measurement, Theory

Concepts, Extraction, Ranking, Selection

## 1. INTRODUCTION

The way we gather, represent, and index the web is changing fundamentally to allow a more semantic view of content. The state of the art in indexing and searching the web is essentially based on viewing it as a collection of hyper-linked pages (each containing a bag of words). The value of the web, however, lies in the wealth of information provided by these pages on a broad range of entities, events, and topics

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

PODS'09, June 29–July 2, 2009, Providence, Rhode Island, USA.  
Copyright 2009 ACM 978-1-60558-553-6 /09/06 ...\$5.00.

(which we collectively refer to as *concepts*). We believe that extracting and organizing metadata in a concept-centric way that allows us to retrieve and present the most relevant information for the concepts related to a user's information need is the next transformative step in the evolution of the web. This will not only improve how current search engines identify and rank relevant content but also allow us to support richer attribute-oriented search criteria and to produce results that are meaningful concept-centric syntheses of information scattered across multiple web pages.

In this paper we outline a research agenda towards this ambitious goal, drawing upon many research threads from the database, information retrieval, machine learning, and web search communities. We begin with a discussion of what we mean by concepts and some of the challenges in representing and organizing concepts and instances of concepts in Section 2. We then motivate our proposal for developing a web of concepts using a combination of example scenarios and data from Yahoo! logs in Section 3. We discuss the challenges in extracting and reconciling concept-centric information from the web in Section 4. A central point of this paper is that we need to develop domain-centric extraction and maintenance systems in order to enable a web of concepts, going beyond the state of the art in site-centric wrappers, domain-independent extraction, and the development of techniques that address a part (e.g., CRF-based attribute extraction algorithms) of the problem but not the whole end-to-end challenge. In Section 5 we discuss several applications, including novel web search paradigms, that are enabled by a web of concepts. We discuss some related work in Section 6 and outline several research challenges in Section 7.

## 2. CONCEPTS

A discussion of entities, concepts, identities, and objects quickly enters deep epistemological and ontological waters. In this paper our focus is on a pragmatic approach to organizing information gathered from the web in a concept-centric fashion; thus, we want to examine how to describe and represent concepts (whatever they are!) in a way that is amenable to indexing and retrieval. To this end, we will avoid rigorous definitions of central terms such as *concepts* and *objects*, and offer informal discussion and examples instead. We will, however, present one specific proposal for representing and storing concept-centric metadata, and the challenges associated with it, in order to make our discussion more concrete.

## 2.1 What are Concepts and Objects?

We use the term *concept* to refer to things of interest to users of the web who are either searching for information (e.g., is there a good Indo-Chinese restaurant in San Jose?) or trying to accomplish some task (e.g., find a good restaurant near downtown and make dinner reservations). Common types of concepts include entities, events, and topics.

Examples of concepts, in our sense of the term, include *restaurants*, *events*, and *academia*. Consider the concept of restaurants. Instances, of course, are specific restaurants, and each is described by *attributes* such as location and cuisine type. It is natural to think of a restaurant as an *object*, with a unique identifier (perhaps its address), state, etc. However, not all concepts have this property. Consider *events*, such as such as concerts and sports; instances include specific events such as the heavyweight title match on Jan 20, 2010 or the 2010 NFL Superbowl between the Packers and the Steelers. Sites such as provide extensive event-centric data. One could associate an id with an event, but are events concepts? To take the discussion one step further, consider the concept of *academia*. We all understand what this denotes, and appreciate that there are many facets to it. We may want to get more information on some aspect of it (e.g., models of tenure in the US versus Europe), but there is no underlying object that corresponds to the topic *academia*. Nonetheless, when we focus on a specific facet, say research publications, we may identify related concepts such as *publications* and *research institutions*, each with instances that are concepts, described by one or more attributes.

## 2.2 Representing a Concept

One of our goals is to enable an evolutionary shift from current search technology, which is based primarily on massively scalable inverted index implementations. We therefore propose to describe an instance of a concept as a *loosely-structured record* (or just *lrec*) that consists of (*attribute-key, value*) pairs. We make two important stipulations:

1. There is a distinguished key called *id*, with the property that it uniquely identifies the record in the stored corpus. However, we may discover that two distinct records in fact describe the same real-world concept, or that a record conflates information about two distinct real-world concepts. This is part of the challenge in determining what information on the web pertains to a given concept and to a given instance of a concept, and is closely related to the *entity matching* problem in the literature; we will discuss this in Section 6.
2. For each concept that is represented in our corpus, we have metadata, including such things as a listing of attributes for which we might have values for one or more instances of that concept. Given a record, we can determine the corresponding concept.

Thus, a concept can be loosely thought of as a *type*, in that we typically store data about several instances, all of which have something in common based on the nature of the underlying concept. In concrete terms, several instances of a concept are likely to have values defined for each of a set of attributes. However, we do not assume that every concept instance defines values for all attributes. Indeed, the set of attributes for which an instance has defined values may

evolve, and the set of attributes associated with a concept may also evolve.

This is a minimalist representational model. In practice, we may need to augment it with additional details at both the record and the concept levels. This model, however, gives us considerable flexibility in dealing with issues characteristic of the web, including missing data, uncertainty about many aspects of the data, and constant evolution (see Section 7.3).

We close this section by introducing one additional notion. A *domain* is a set of related concepts. Thus, people, publications and conferences are examples of concepts in the academic community domain.

## 2.3 Concept-Centric Data Organization

Undertaking to organize all the information on the web at a semantically useful level is no mean task, and that is essentially what is involved here. We make no claim to having a comprehensive solution, but outline some important issues to be addressed. There is a considerable literature on knowledge representation [13], to guide any detailed discussion of representational issues. For concreteness in our discussion, however, we have deliberately separated what we believe to be a minimal core for representing semantic, concept-centric descriptions in a manner that is amenable to leveraging existing search engine infrastructure; this is the basic lrec representation we presented above. Almost certainly, it will be worth considering extensions to this core, and the question is which extensions make sense and what they additionally support. In this context, some nuances worth considering include the following:

- Should the basic model allow for nested structure, e.g., XML-style paths and nested references to lrec ids, rather than just a flat collection of attribute-value pairs in an lrec? We have suggested a simpler alternative in part because retrieval is more readily mapped to existing inverted indexes, but also because populating more sophisticated representational structures using extraction is likely to be harder.
- Should there be support for provenance, versions, and uncertainty? It is by now traditional for large search engines to include information directly on the search results page providing metadata and links for an object of interest. If the query references a restaurant, for instance, there will be a box with a map, contact information, ratings, and so forth. Connecting users to key pieces of text (professional reviews, user reviews, blog mentions, etc.) is a critical part of building concept-oriented search engines. Furthermore, a piece of text may reference multiple concepts, the association to concepts is inferred and therefore uncertain, and may change over time, so simply associating the text with a single object is not sufficient: we will need to identify and maintain references between concepts of different types, and perhaps maintain versions of important concept instances over windows of time. These are computationally challenging requirements. How best can we approximate what is essential with an acceptable cost?
- In the real world, concepts are interrelated by many natural taxonomies and containment or inheritance re-

relationships. How far should we extend support for organizing lrecs into corresponding hierarchical relationships? As examples of the nuances we may want to capture, consider the following:

- We may want to be able to talk about the Nikon D40 model of digital cameras; how it is a particular kind of digital camera, which in turn is a particular kind of camera; and how it is a kind of Nikon cameras.
- We may want to be able to talk about the D40 camera as a concept, with each physical unit of this type as an instance (e.g., in the Amazon inventory); we may also want to talk about the D40 camera as an instance of the concept of camera models (e.g., in the review forum).
- We may want to talk about how the D40 camera *is part of* a special camera package.
- We may want to talk about concepts such as academia or war that have many facets, and are hard to categorize unambiguously into given taxonomies. Yet, a collection of such concepts may lend itself to hierarchical categorization techniques that yield a data-driven taxonomy. What is the relative role of the two approaches—categorizing concepts and instances into curator-developed taxonomies versus data-driven taxonomy construction?

The distinctions we have drawn are only illustrative; clearly, there are more representational issues to be considered. Beyond issues of expressiveness and computational cost, however, in our setting there are two new considerations. First, is it feasible to extract or otherwise obtain (e.g., via contractual feeds) information about concepts and instances at this level of detail, reflecting such taxonomic niceties? Second, can our ability to interpret what users are looking for and to match it with the concept database take such nuances into account reliably? What is the right granularity for a concept hierarchy, given our ability to automatically categorize and match to user intent?

### 3. USAGE STUDIES

In this section we consider the implicit role of concepts in online user activities such as web search and browsing. We envision the following scenarios:

- A user may have a specific instance of a concept in mind and wants to search or browse for various attributes associated with that instance, e.g., find the menu, or reviews, for a given restaurant.
- A user may want to search for a set of concept instances whose values satisfy certain conditions, e.g., the best bakeries in Cupertino, or the closest restaurant to a given theater.

By examining users' search/browsing behavior, we hope to gain some insights into their activities with respect to concepts. Notice that in the search setting, users may also benefit from aggregated information, e.g., information on a specific concept hosted by different websites.

We present a set of empirical observations based on analyzing Yahoo! Search and Yahoo! Toolbar logs, focusing on how the notion of concepts may be relevant to user online experience and how they may benefit from a concept-oriented interpretation of the web content on the part of search engines.

**Concepts vs. Search.** The question we seek to answer is if users indeed search for both a specific instance of a concept and a set of concept instances with values for some keys satisfying certain constraints. To do this, we looked at queries (over a month) resulting in a click on a URL from , a site hosting reviews on various local businesses. We identified three main sub-categories of these URLs:

- 59% are biz URLs, where each page is about an individual business,
- 19% are search URLs, where each page is a search result page in , which can be searching for either a specific business (e.g., a restaurant named churro factory in Chicago) or a group of businesses (e.g., wedding cakes Los Angeles).
- 11% are c URLs, or category URLs, where each page is about a group of businesses in a pre-defined category, e.g., San Jose Italian Restaurants.

Thus, at least for searches that result in clicks on this website, we estimate that people search for a specific instance roughly 60% – 70% of the time and search for a set of instances roughly 10% – 20% of the time. Even if these specific numbers might vary for other websites, our study provides some evidence that users do conduct significant amounts of both types of search.

**Searching for Attributes of a Concept.** We now proceed to ask if users explicitly search for different attributes of a concept. To study this, we performed experiments on the search logs. First we obtained a list of restaurant homepage URLs from . We then examined queries that led to a click on one of these restaurant homepage URLs. After removing the restaurant names and location information from the queries, we tallied the remaining tokens. Top words that co-occur in conjunction with restaurant names are the following: menu (3%), coupons (1.8%), online, weekly specials, locations (1.5%), etc., where the numbers in the parenthesis show the fraction of occurrence. Note that these are queries that led to a click on the restaurant's homepage, even when the user was actually looking for a specific attribute. The fraction of co-occurring words remain largely the same even when the query merely contained the word restaurant, regardless of what the user clicked, or when the query contained more specific terms, say, chinese restaurant. Other interesting but less frequently occurring attributes that surface from our study include nutrition, to go, delivery, careers, and cod.

**Value in Aggregation.** Would users benefit from aggregated information? If we look at users who clicked on a biz type URL from Yelp, more than 59% of the time they also clicked on at least one other URL for the same query, and 35% of the time they clicked on at least two other URLs. Manual examination of a small sample of the queries that

led to these URLs finds most of them to be indeed looking for a specific business and other pages clicked can include the homepage of the business, profile pages from other aggregation sites such as , , etc., as well as blogs and reviews written about the business. Clearly, even when users are searching for a specific instance of a concept, they seek diverse information beyond one source. An aggregation page can facilitate this process (Section 5.2).

As a concrete example of a more sophisticated search, here is a query from the log: mexican food chicago best salsa and the user clicked on a Yelp category page, a Citysearch search result page, as well as a page of expert reviews hosted by a restaurant. Most likely the user was consulting multiple sources to reach a conclusion regarding the best salsa. An aggregated page with locations of different mexican food places in chicago, accompanied by reviews that commented on salsa from different sources, with meta information on the trust-worthiness of these sources could probably add more value to this user's search for the best salsa.

**Concepts vs. Browsing.** How do users get to concepts on the web? To understand this, we analyzed the toolbar logs. As before, we obtained a list of restaurant homepages from ; these homepages are meant to represent the instances of the restaurant concept. We then examined the user trails that pass through one of these restaurant homepages. First, we observed that about 42% of the homepage visits are immediately preceded by a query to a search engine. Next, we examined the URL that is surfed immediately after the homepage is visited and observed that about 11.5% of these URLs are the location/address of the restaurant, 9% of them are the menu pages, and about 1% of them are coupons. These studies demonstrate that users are already looking for natural attributes of the restaurant concept, even in a browsing mode. Furthermore, about 10.5% of the user trails contain more than one distinct instance of the restaurant concept, suggesting that an aggregation of instances might be beneficial to the user.

## 4. CONSTRUCTING A WEB OF CONCEPTS

We can view today's web as a simplified web of concepts, where each record is of type "Document." We want to start from here and extract records of richer types. We use the term *extraction* broadly to refer to any of the operations that either create new records belonging to the concept or enrich existing records. The following list of extraction operations is typical.

- **Information extraction.** This extracts structured data from documents, e.g., an address from a restaurant website or a list of publications from a personal homepage.
- **Linking.** This creates links between existing records. Examples include linking a restaurant review to its corresponding restaurant or matching two different mentions of the same author.
- **Analysis.** This attaches metadata to records, e.g., identify and tag the cuisine type for a restaurant website or assign sentiments to a review.

We propose constructing a web of concepts using *domain-centric* extraction. A domain specifies a set of concepts of

interest that may not be specific to any individual source. For example, a restaurant domain might specify the concepts menu, location, review, an academic domain might specify author, publication, and a shopping domain might specify product, seller, review. Most of the current extraction methods are *site-centric*, i.e., they can only be deployed to extract from a specific website or data source. In contrast, to construct a web of concepts, we propose to look at extraction holistically on a domain level. To do this, we need to bridge the gap between the capabilities of existing methods and the requirements for domain-centric extraction.

We start by describing the current extraction methods and then describe the domain-centric extraction methods that we are currently pursuing.

### 4.1 Site-Centric Extraction

The existing site-centric (or source-centric) methods for extraction can be classified along two dimensions: structural and semantic.

The *structural* approaches typically rely on the rich HTML structure employed by the author for presenting the content. For example, tables and lists are frequently used in web-pages to present structured content. Restaurant aggregators (e.g., , ), product catalogs (e.g., , ), music and movie websites (e.g., , ), and virtually any website that serves pages off a database often uses scripts to generate highly structured and regular HTML, and thus impart structure across multiple pages within the website. In this direction, *wrapper induction* [22, 44, 49, 40, 21, 39, 58, 7, 50, 4] has been used as a powerful and robust way for structural site-centric extraction. With relatively few labeled examples, extraction rules, called *wrappers*, can be learnt to extract information from a specific website. The main drawback with wrappers is that they rely on the existence of a structure.

In contrast, *semantic* approaches view web pages as text documents and employ natural language processing [16] and machine-learned probabilistic models. For instance, Conditional Random Fields [33, 36, 46] have been used effectively to parse postal addresses and lists of publications. Semantic techniques often require large supervised training data, and are sensitive to the construction of this training data; e.g., a probabilistic model learnt to extract US restaurant addresses may completely fail on European restaurant addresses, and a model learnt to extract Computer Science publications may perform poorly on Physics publications.

### 4.2 Domain-Centric Extraction

The main challenge in constructing a web of concepts is to extrapolate site-level extraction techniques to work at the domain level. A domain has an unbounded number of websites, which renders pure wrapper-style extraction infeasible, and a large amount of diversity, which cannot be easily captured by probabilistic models. While a complete solution towards constructing a web of concepts requires several challenges to be solved (Section 7.2), we describe some of the ongoing work at Yahoo! Research, and briefly discuss work that is especially relevant from other research groups.

**Domain-Centric List Extraction.** The goal of this work [41] at Yahoo! Research is a domain-centric extraction of lists. The motivation is that much information on the web is present in the form of lists or tables, e.g., a list of restaur-

rants, a list of menu items for a restaurant, a list of publications, etc. This work tries to extract content in the form of lists by combining structure and semantics. A list can often be identified on a webpage by a repeating pattern of HTML structure. However, webpages often contain several lists, and we need to identify the lists that we are interested in; this typically requires us to combine domain knowledge with structural cues. For example, to extract a list of restaurants, we might have two kinds of domain knowledge: first, the fields of interest such as the address, city, zip code, phone number, and hours of operation, along with rules to identify zips/phones, and second, certain statistical properties (e.g., each restaurant is associated with a single zip code and has one or two phone numbers). This domain knowledge, along with an analysis of the repeated structure, is used to identify lists of addresses in a completely unsupervised, site-independent fashion.

**Relational Classification.** Most large web sites have pages covering a variety of content. One might be interested in filtering out only those pages that belong to a certain category and then doing further extraction on them. For example a city site such as [www.chicago.gov](http://www.chicago.gov) has pages on diverse categories such as hotels, attractions, night-life, restaurants, events, etc. Let us say we are interested in selecting only events pages and extracting individual event information from them. If we are also interested in repeating the same process on thousands of city sites, then it is not scalable to develop an events classifier for each site since it requires labeling several pages on each site as events/non-events. Developing a global events classifier is easier, but it tends to be noisy given the vastly different content in the large collection of sites. Fortunately, the link and directory relationships in a site contain valuable signals for solving such a classification problem. For instance, all the events pages in [www.chicago.gov](http://www.chicago.gov) are placed in a directory called calendar. Note that this relational structure will be different for different web sites. After bootstrapping the pages of a site with the classification labels given by an inaccurate classifier (such as the global classifier mentioned above), the relational structure present in that site can be used to revise them and get highly accurate classification. Graph-based methods for doing the above are discussed in [60].

**Aggregator Mining.** Independently maintained concept-centric web corpora will emerge and grow in importance. Many examples already exist: Wikipedia and numerous aggregator sites such as Yelp; sites such as LinkedIn that gather and selectively expose structured information about individuals; and structured integrated corpora such as *Freebase* and *KnowItAll* [30], and community driven *mass collaboration* sites such as *DBLife* [25, 48]. Automatically discovering such concept-centric web corpora, extracting data from them, and integrating with data from the rest of the web can significantly increase the quality of the unified web of concepts that we seek to build.

The aggregator mining work currently in progress at Yahoo! Research aims to address this problem by applying *bootstrapping* to make extraction scalable. The main idea is to use already extracted records to automatically generate labeled data and use it to extract more records. For instance, suppose we have already extracted a small set of Italian restaurant menu items and stored them in a database. Con-

sider a restaurant website that lists a menu. As discussed before, we can identify lists by a structural analysis of the HTML content; this will yield us the menu items. Now, if we can map a few of the menu items to our database (that reflects our current knowledge of Italian restaurant menu items), then we can infer that the list represents an Italian restaurant menu and can extract additional menu items from the list to add to the database. Thus, we can start from a small set of seed records and bootstrap to extract more records from sources that overlap with the current set or records.

**Matching.** The matching work done at Yahoo! Research [23] aims to bridge the structured and unstructured views of a record by establishing that a piece of text is “about” a record. For example, matching a piece of text that is known to be a restaurant review to its corresponding restaurant will fall in this category. The main idea is to develop a domain-centric generative model of text that takes into account the domain knowledge (e.g., address, city, cuisine, etc. for restaurants) and use this model in order to choose the most likely restaurant a given review is about. This work is related to the widely studied entity matching problem, which we discuss in Section 6.

**Other Work.** The *DBLife* project at Wisconsin [25] shares the goal of building concept-centric integrated web repositories, and has led to the creation of [dblife.cs.wisc.edu](http://dblife.cs.wisc.edu) and influenced our work at Yahoo!. In contrast to our strong emphasis on domain-centric training of machine learned models, however, the *DBLife* project has taken a more rule-oriented approach and emphasized issues such as mass collaboration, managing site evolution, and optimizing extraction. The *Avatar* project at IBM Almaden [42] aims to apply extraction to a diverse range of datasets to improve search over enterprise datasets. It also takes a rule-oriented approach, and emphasizes rapid rule development and optimization. The *WebTables* work of Cafarella et al. [15] aims to extract all tables from the web and store them as relational data. The techniques are domain-centric. The main challenge here is to attach semantics to the extracted data for it to be useful in populating a web of concepts. An orthogonal approach is *DeepWeb* [45], which aims to surface the deep web by automated form-filling techniques. The *KnowItAll* [30] and the *TextRunner* work at the University of Washington also aim to extract information from the web in a site-independent manner using simple yet effective natural language techniques. The focus however is on extracting “common knowledge”, i.e., the facts described textually in natural language and that occur often at several places on the web.

## 5. APPLICATIONS

In this section we discuss a variety of applications that are enabled by a web of concepts.

### 5.1 Augmenting Web Search

We begin with applications in the domain of web search. First, it is by now traditional for large search engines to include information directly on the search results page providing metadata and links for an object of interest. If the query references a restaurant, for instance, there will be a box with a map, contact information, ratings, and so forth. If

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/898070140101006050>