The Web Data Factory

Search, Integrate, Organize… The Real World

The Web has evolved from its early days as a big *library* of interconnected "hypertext documents" to its modern form as a giant *database* of rich "structured data." Increasingly, we rely on the Web for a variety of information needs, ranging from:

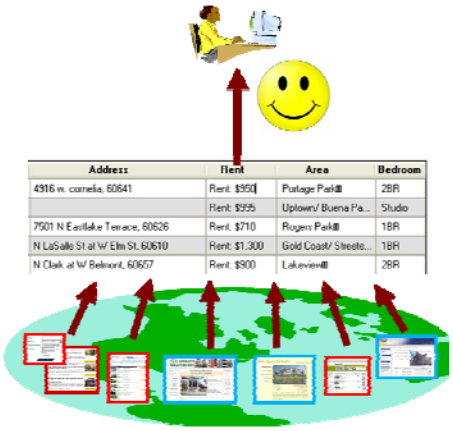
- Our needs in our personal lives (e.g., when looking for rental housing, cars to buy, restaurants to eat, places to travel, available jobs, doctors to consult, research grants to apply for, and schools to attend), to
- Successful business operations (such as targeted behavioral marketing, finding the best suppliers), to
- Intelligence research (e.g., brand sentiment analysis, consumer spending trends, market research, and more).

The Problem Today's search engines simply index the keyword content of Web pages, lacking the ability to understand the semantics of the data present those pages. In the absence of a "data-aware" search engine, our information discovery is often tedious and inefficient—beginning from the general search engines to find relevant websites, and then manually browsing through each of these websites to find information matching our specialized search criteria.

- Consider, for example, a college student in her freshmen year who is looking for an apartment for rent. Can she use keyword-based search engines to find *"all **1- bedroom** apartments **within 3 miles** of the Stanford campus, preferably **furnished**, with a monthly rent of **less than $1000**"?* She will need to browse through the listings on many different sites—starting from large Internet listing sites, to local newspaper classifieds, and finally to a large number of individual landlord websites.
- Likewise, consider an intelligence officer in the U.S. Army who is assigned the task of constructing a database of every mosque in Afghanistan, including the **name** and **geo-coordinates**. There is no single Web source that can provide a complete listing—the officer would need to extract and aggregate this information from many different online directories, organizational sites, travel portals, or forums.

Cazoodle is a startup company originating at the University of Illinois at Urbana-Champaign, which developed large-scale information integration technologies—the Web Data Factory—to produce actionable data from unstructured HTML content in both the surface and the deep Web. The company was founded in 2006 by Professor Kevin Chang and his research team, who together developed innovative and unparalleled deep Web search technologies.

Technology innovation Cazoodle has built an innovative data factory solution for large-scale data integration on the Web. This intelligent data gathering technology can easily crawl, extract, and integrate data from thousands of Web sources in a structured format, e.g., crawl all the rental listings from an apartment listing website and index detailed attributes such as the number of bedrooms, number of bathrooms, rental cost, address, and images. Our solution is capable of understanding the semantics of the data in any domain, and thus, can enable novel search engines in a variety of application domains. We are using this technology to serve both the enterprise customers in their data crawling needs, as well as the end users by providing them novel search engines.



Key competitive strength Our data factory solution has many benefits over the state-of-the-art solutions for data crawling. The problem of transforming unstructured text content to a structured database format has long been studied in both industry and academia. However, none of the existing solutions satisfies all the desired criteria, as explained in the following comparison chart:

Different Categories of Solutions →	**Manual Copy-paste**	**Software Scripts/ Program**	**Automatic Wrapper Induction**	**GUI for writing scripts**	**Cazoodle Data Factory**
Illustrative Examples	Data entry jobs	in-house/ off-shore software engineers	Road Runner, HLRT, etc.	Kapow, Fetch, Dapper, etc.	A new solution
Evaluation Results *Based on our personal judgment. Customers should use their own due diligence when making their choice.*	✓ = Satisfies the criteria ✗✗ = Represents the degree of dissatisfaction				
1. Accuracy *Is the extraction quality always correct for all the records for any new application and website?*	✓	✗✗	✗✗✗✗	✓	✓
2. Scalability *Is it proven to handle thousands of websites including dynamic deep Web interactions and foreign language?*	✗✗✗ ✗✗✗	✗✗	✗✗✗✗	✗	✓
3. Complexity *Can a person with no Web programming skills configure new applications and websites?*	✓	✗✗✗	✓	✗✗	✓
4. Affordability *Are the costs of maintenance and computation low, to support repeated crawling?*	✗✗✗ ✗✗✗	✗✗✗	✓	✗	✓

Our mission is to enable novel "data-aware" search engines in a variety of application domains. Unlike the general search engines that simply index the keywords in the Web pages, our technology can index the detailed structured attributes of the data in these pages.

Product Directions We are building our products and services along the following two directions:

I. **Data Factory Service**: On one hand, we are offering our data factory solution as an end to end service to enterprise customers to power up their backend data gathering needs.

II. **Vertical Search Applications**: On the other hand, we are building new consumer applications in selected domains to be used directly by the end users.

Product Direction I. Data Factory Solution for Enterprise Customers

We provide an end-to-end solution for supporting the backend data crawling needs of enterprise customers. In many scenarios companies can benefit from access to large amounts of structured data, aggregated from a variety of different Web sources. We offer our services for real time tracking (e.g., to track the availability of products in local stores), as well as for large-scale offline crawling.

- ✓ 100% accuracy in extraction,
- ✓ Understands the semantics of data in any application domain,
- ✓ Scales to thousands of online Web sources, and
- ✓ Affordable to build a viable business.

Customers may use our data factory solution in a variety of situations, as illustrated in the following examples:

Case Study I

Need for crawling and indexing rich content to support an online retail business.

Problem:

A major online retailer in the United States wants to expand into selling wines online. Today, the retail wine industry is estimated at $27 billion per year, but only 10% of the sales occur online. The wine retailer can set up partnerships with existing suppliers to build inventories; however, the suppliers provide only basic information about wines, e.g., wine name, alcohol level, and pricing. It is imperative that the online retail website must contain rich information about the wine, including tasting notes, varietal information, vintage, appellation, critic reviews, user reviews, etc. Such rich information is distributed over thousands of winery websites, e.g., www.bridlewoodwinery.com and www.raymondvineyards.com.

Solution:

The wine retailer can use Cazoodle's Data Factory service to power up their backend data crawling to index all the rich content from all winery websites, in a consistent schema. As an illustration, the Data Factory service would return the following xml files with all the records from the two example websites www.bridlewoodwinery.com and www.raymondvineyards.com:

- ➢ http://www.cazoodle.com/partners/example/bridlewoodwinery.xml and
- ➢ http://www.cazoodle.com/partners/example/raymondvineyards.xml

Case Study II Need to gather local events information for an entertainment website.

Problem:

A large online entertainment website provides information about many national events such as new movie releases, major festivals, sports, large conferences, etc. The website is facing immense challenge from a few local community websites that are tailored towards specific geography. To improve service for its audience, the entertainment website must aggregate the events that take place in small local communities.

Solution:

The Cazoodle Data Factory service can identify the local newspaper website, event venues, art theaters, organizational calendars, etc. and crawl and index highly local events nationwide. The technology is already proven to scale to thousands of online sources in other application domains (i.e., real estate for rent, online ecommerce, and vacation rentals), and so local events would be just another application. For example, consider the following structured data obtained from a sample of events websites:

- http://www.cazoodle.com/partners/example/jazz88.xml
- http://www.cazoodle.com/partners/example/iloveny.xml

Case Study III Need for aggregating geo-spatial features for Army intelligence databases.

Problem:

The U.S. Army is in need of a technology for building comprehensive and up to date databases with intelligence information. While much of the information is available online, the current approach of maintaining the databases is way too expensive—soldiers or intelligence expert would need to manually collect information by visiting individual website.

Solution:

The Data Factory solution can quickly identify all Web sources relevant for the specific data needed. Consider for example, the task of compiling a database of all the mosques in Afghanistan, including their variant names, and geo-coordinates. The Data Factory solution can quickly identify the relevant websites including religious organizational websites, travel blogs, community forums, etc., and then crawl all the content to generate a structured database.

Product Direction II. Novel Search Applications for End Users In addition to providing our Data Factory solutions to enterprise customers, we are also building novel search applications that can directly benefit end users. Our applications showcase an unparalleled scale of Web data integration, both in the number of websites that we crawl, and in the number of records we index. Our current applications for end users include the following services:

1. Apartment Rentals (http://apartments.cazoodle.com)

Over 15% of American households currently live in rental dwellings. Of all new home searches, over 80% begin online; and 96% of renters have to browse through listings on multiple websites. By providing a one-stop search for rental listings that are aggregated from more than 10,000 apartment rental websites, Cazoodle is changing the way renters search for apartments online,.

2. Electronics Shopping (http://shopping.cazoodle.com)

Shopping is booming online, with 2008 ecommerce at $156B (Forrester), of which 40% is for electronics products. Most comparison shopping sites rely on merchants to submit data feeds—and as a result, their coverage is often limited. Cazoodle organically crawls and integrates more than two million product offers across 15 consumer electronics categories—thus transforming the Internet into one organically integrated marketplace.

3. Vacation Rentals (http://vacation.cazoodle.com)

Family reunions and leisurely travelers are increasingly preferring vacation rental option rather than reserving multiple adjacent hotel rooms. Over 2.45 million family reunion travelers indicated staying in a vacation rental in 2008. Cazoodle organically crawls the vacation rental listings from over 4000 vacation rental listing websites—making it so easy for you to discover your perfect home away from home.

4. Geospatial Search (http://geoengine.cazoodle.com)

In collaboration with the U.S. Army (under a SBIR Phase II grant), we are building technologies for generating intelligence databases by aggregating the information publicly available on the Web, albeit distributed across a variety of Web sources. Applications include enriching existing geospatial databases, as well as finding new geospatial entities to support the Army's intelligence database generation task force.

Business Model

We have three channels of revenue generation.

1. Lead-generation For our search products, we reserve a section at the top of the results for the advertised featured listings. Our structured search products inherently allow us to provide a far more targeted advertising, e.g., the user queries on our apartment rentals products include the geography of interest and renter's budget.

2. Data gathering partnership Our Data Factory solution generates revenue in the form of data gathering fees from partner companies by providing them with a white-labeled data gathering ability. Current partners include the University of Illinois, a major retailer in U.S., and a major international media group.

3. Intelligence market research Our ability to collect a large volume of data continuously over time provides a great data set for intelligent marketing research. For example, a) The data archive we generate regarding apartment rentals can provide great insights into rental trends all over the U.S.; b) The up-to-minute price tracking ability of our shopping search product can be an invaluable tool for competitive price tracking.

Position in Search Industry

There are many search engines already prevalent on the market. How are we unique? Our technologies can *interact* with the structured data hidden behind query forms of dynamic "deep Web" sites—beyond static links reachable by current engines. Would users find our service more useful than what they are now using—Shopping.com or more basically, say, than Google Search? Absolutely! We offer a unique value proposition that builds on and complements all the existing services in a way that is totally unmatched by any existing service.

1. General search engines such as Google, Bing, and Yahoo are great at finding useful Web sources. Often, these results are only the starting point for information discovery. Our services go deeper to extract the structured attributes of data on those Web sources--there by allowing users to easily filter the results through their desired query conditions, and visit only the sources that provide data that matches their needs.

2. Internet listing services are popular in many domains, e.g., in the apartment rental domain, we have rent.com, apartments.com, craigslist.org, Google Base, etc. Likewise, in the shopping search domain, we have shopping.com, pricegrabber.com, and Google products. All these services rely on service providers to submit (or advertise) their XML feeds, and thus, are inherently limited in their coverage. Our products integrate data from existing listing services, as well as directly from Web sources via organic crawling.

3. Semantic search engines are emerging to transform the raw text of the Web into an intelligent knowledge base, using advanced natural language parsing, e.g., powerset.com. While our solutions are also based on machine learning techniques, we are not aiming to mine patterns or hidden semantics in raw text. Our key capability is in our Data Factory technology, which transforms unstructured text into a structured format.

4. Vertical search engines already exist in specific topical domains, e.g., righthealth.com in the health domain. Such applications work well in providing expertise from a few selected sources. Our technology, however, can further enrich their usability by tapping into the wealth of information distributed in heterogeneous web sources.

For more information:

Govind Kabra, CTO, Cazoodle, Inc.
60 Hazelwood Drive, Suite 122
Champaign, IL 61820-7460

govind.kabra@cazoodle.com
Office 217-864-8378
Mobile 217-419-2637