

Deep crawling. Attribute parsing. Universal search.

Google Presentation

BytePlay

November 26, 2007

- 1 Motivation
- 2 Crawling
- 3 Parsing
- 4 Search
- 5 UI Demo & Questions

Attribute-relevant, deep-crawl results

The screenshot shows a web browser window with the address bar containing 'http://www.byteplay.com/google/'. The search bar contains 'chelsea 2 beds'. The search results are displayed under the 'Property' tab, showing 1-10 of about 1,740,000 results in 0.10 seconds. The first result is 'Property search results for chelsea 2 beds', which includes a thumbnail image of a bedroom and several listings with prices and agent names. The second result is 'Property for Sale in Chelsea 2 bedrooms', and the third is 'London Accommodation - 2 Bedroom Rental in Notting Hill ...'. On the right side, there are 'Sponsored Links' for 'Chelsea FC - DVD', 'Huge Savings Mattress', 'Mega Bed Sale Now On', and 'Buy Beds At Argos'.

extate ▼


extate: the property search en... chelsea 2 beds – Google Search

Web Images Video News Maps Mail more ▼ | My Notebooks | Web History | My Account | Sign out

Google chelsea 2 beds Search Advanced Search Preferences

Web Property Results 1 - 10 of about 1,740,000 for [chelsea 2 beds](#). (0.10 seconds)

[Property search results for chelsea 2 beds](#)

 [Chelsea Embankment, SW3, London](#) - £725,000 - John D Wood & Co
This is a charming 1/2 bedroom flat on the 3rd floor of a handsome red brick, period building with fabulous southerly views across the River Thames ...
[See Map](#)

[Chelsea Manor Street, SW3, London](#) - £595,000 - Carter Jonas
[Beaufort Street, SW3, London](#) - £825,000 - Strutt & Parker

[See chelsea 2 beds results, available through Google Property »](#)

[Property for Sale in Chelsea 2 bedrooms](#)
Find all flats to buy in **Chelsea 2 bedrooms** and browse each flat with a **Chelsea** map and local information.
[www.nestoria.co.uk/chelsea/flat/buy/bedrooms-2 - 102k - Cached - Similar pages - Note this](#)

[London Accommodation - 2 Bedroom Rental in Notting Hill ...](#)
apartment type, **2 Bedrooms**. neighborhood, Kensington - **Chelsea** - Notting Hill - London.
address, Cambridge Gardens between Portobello Road & Ladbroke Grove ...
[www.nyhabitat.com/london-apartment/vacation/44 - 39k - Cached - Similar pages - Note this](#)

Sponsored Links

[Chelsea FC - DVD](#)
Behind the scenes feature length film "Blue Revolution" Out Now!
[www.amazon.co.uk](#)

[Huge Savings Mattress](#)
Thousands of Mattress in Stock With Free Next Day Delivery UK
[MattressMan.co.uk](#)

[Mega Bed Sale Now On](#)
Massive Savings Upto 40% Off Many In Stock For Fast Delivery
[www.bedstar.co.uk/beds](#) England

[Buy Beds At Argos](#)

Deep crawling

- Beyond vanilla hyperlinks:
 - forms, drop-downs
 - cookies
 - URL rewrite rules
 - JavaScript links
 - generating permalinks

Deep crawling

- Beyond vanilla hyperlinks:
 - forms, drop-downs
 - cookies
 - URL rewrite rules
 - JavaScript links
 - generating permalinks
- Server-side state
 - Session time-outs (replay action)
 - Different content at the same URL
 - Same content at different URLs

Deep crawling

- Beyond vanilla hyperlinks:
 - forms, drop-downs
 - cookies
 - URL rewrite rules
 - JavaScript links
 - generating permalinks
- Server-side state
 - Session time-outs (replay action)
 - Different content at the same URL
 - Same content at different URLs
- Crawling strategies
 - limits on number of results returned
 - optimising crawl paths
 - crawling rates: new properties versus modified versus removed

Heuristic framework to parse out attributes

- 1 Comparative variance of document nodes
 - Matched to comparative variance of property attributes

Heuristic framework to parse out attributes

- 1 Comparative variance of document nodes
 - Matched to comparative variance of property attributes
- 2 Unchanging node contents
 - Determining static content—allowing for structural weakness

Heuristic framework to parse out attributes

- 1 Comparative variance of document nodes
 - Matched to comparative variance of property attributes
- 2 Unchanging node contents
 - Determining static content—allowing for structural weakness
- 3 Text classifier
 - Property summaries

Heuristic framework to parse out attributes

- 1 Comparative variance of document nodes
 - Matched to comparative variance of property attributes
- 2 Unchanging node contents
 - Determining static content—allowing for structural weakness
- 3 Text classifier
 - Property summaries
- 4 Attribute consistency
 - Latitude & longitude, postcode, address string
 - Price, floor size, property type, location

Heuristic framework to parse out attributes

- 1 Comparative variance of document nodes
 - Matched to comparative variance of property attributes
- 2 Unchanging node contents
 - Determining static content—allowing for structural weakness
- 3 Text classifier
 - Property summaries
- 4 Attribute consistency
 - Latitude & longitude, postcode, address string
 - Price, floor size, property type, location
- 5 Part-of-speech tagging
 - Quantitative attributes in descriptions

Heuristic framework to parse out attributes

- 1 Comparative variance of document nodes
 - Matched to comparative variance of property attributes
- 2 Unchanging node contents
 - Determining static content—allowing for structural weakness
- 3 Text classifier
 - Property summaries
- 4 Attribute consistency
 - Latitude & longitude, postcode, address string
 - Price, floor size, property type, location
- 5 Part-of-speech tagging
 - Quantitative attributes in descriptions
- 6 Pattern matching
 - Quantitative attributes & postcodes

Framework prerequisites

- Custom DSL for describing nodes and node relations
 - Inspired by XPath language
 - Allows us to query for
 - statistics (node variance, probabilities)
 - document relationships (map page, floor plan, etc.)
 - meta data (URIs, crawl paths, etc.)
 - Collections of heuristics given as DSL expressions

Framework prerequisites

- Custom DSL for describing nodes and node relations
 - Inspired by XPath language
 - Allows us to query for
 - statistics (node variance, probabilities)
 - document relationships (map page, floor plan, etc.)
 - meta data (URIs, crawl paths, etc.)
 - Collections of heuristics given as DSL expressions
- Automatically generate/learn heuristics from sample documents
 - From a each data source
 - Across multiple data sources

Framework prerequisites

- Custom DSL for describing nodes and node relations
 - Inspired by XPath language
 - Allows us to query for
 - statistics (node variance, probabilities)
 - document relationships (map page, floor plan, etc.)
 - meta data (URIs, crawl paths, etc.)
 - Collections of heuristics given as DSL expressions
- Automatically generate/learn heuristics from sample documents
 - From a each data source
 - Across multiple data sources
- Distributed framework for learning (and parsing)

Recall, relevance, universal search

- Tagging to improve recall
 - location information
 - features (outside space)

Recall, relevance, universal search

- Tagging to improve recall
 - location information
 - features (outside space)
- Custom full-text search
 - Boolean text match
 - ordering & proximity of keywords
 - (partial) tag matching as n-grams
 - quantitative refinement
 - geographical relevancy

Recall, relevance, universal search

- Tagging to improve recall
 - location information
 - features (outside space)
- Custom full-text search
 - Boolean text match
 - ordering & proximity of keywords
 - (partial) tag matching as n-grams
 - quantitative refinement
 - geographical relevancy
- Single search box

Recall, relevance, universal search

- Tagging to improve recall
 - location information
 - features (outside space)
- Custom full-text search
 - Boolean text match
 - ordering & proximity of keywords
 - (partial) tag matching as n-grams
 - quantitative refinement
 - geographical relevancy
- Single search box
- Show map when relevant
 - Query can be disambiguated
 - All properties can be seen without paging

Demo & Questions



e.g. Edgware Road, London or Flat N1, 300-400 pw, balcony or 3-4 bed bungalow, South Yorkshire

 Property Search

 I'm Feeling Wealthy