

# Delivering Query Suggestions from Authority Data in Library Discovery Layers

Ben Pennell, Xi Niu

# What is Query Suggestion anyways?

- Auto-suggest, suggestion service, suggest, etc
- In general:
  - “...a user interface interaction method to progressively search for and filter through text. As the user types text, one or more possible matches for the text are found and immediately presented to the user.” – Wikipedia
- Common feature in most commercial search interfaces:
  - google.com, amazon.com, etc
- A feature so powerful it can suggest your whole life for you.



# Intro to Search UNC

- UNC's Online Catalog (one of two)
- <http://search.lib.unc.edu/>
- "Discovery Layer"
  - Faceted
  - Keyword searching
- Endeca based
  - Commercially licensed, faceted indexing software
- TRLN - Triangle Research Library Network
- Minimal interaction with authority data by default



# Authority data?

- TRLN looking for ways to get authority data to users.
- Hypothesis time
  - Giving users immediate and readily understandable access to authority data at the point of need will lower barriers to usage and provide a better search experience.
- Preferably without users knowing what authority data is.



# An LCSH in every search box

- Approaches:
- Facets - Subject, Author
- Subject linking and broadening
- Subject result suggestions
- Author result suggestions - OCLC NameFinder
- Query Suggestions



# Auto-Suggest profile

- Developed in 2009, released January 2010, adopted by TRLN August 2010
- Solr based
  - What is Solr? Open source, faceted search index.
  - Sounds familiar, why not Endeca?
  - Cut out the unneeded parts, make your own schema
- Populated from MARC Fields
  - Pulled straight out of catalog records
  - Author, subject and title



# Lets look at that "Pulled out of catalog records" bit a little more closely

- Extracted from 5.4gb of pre-filtered text MARC records
- Containing ~36.6million applicable field values
  - 12.5million unique suggestions
- Python scripts - not my first language
  - Data cleanup, truncation
  - Lots of fun with efficiency refactoring
- Loaded nightly



# Several iterations of python loading scripts later...

<u>ac</u>	type	<u>occurs</u>	<u>source</u>	<u>ackey</u>
shakespeare, william, 1564-1616	author	589	unc, filmfinder, duke, nccu, ncsu	shakespeare,_william, _1564-1616 author
adventures of huckleberry finn	title	45	unc, duke	adventures_of _huckleberry_finn title
united states. congress. house -- rules and practice	subject	27	unc, duke	united_states._ _congress._house_-- _rules_and _practice subject
mama cat has three kittens	title	1	unc	mama_cat_has_ three_kittens title





## Now for the meat/tofu of it

- The key purpose of an auto-suggest feature:
  - To try to extrapolate where the user is headed from very little info, and help them actually get there
- Solr is good at this:
  - Built in search relevancy
  - Allows complex formulas for tuning ranking



# Suggestion Querying

- Keyword searching
- Weighting and Tie-breaking
  - Word order matching
  - occurs field
- Stopword heavy queries
  - Full begins with matching
- Single-word versus Multi-word queries
  - Last term (active term)
  - Partial Match or Complete Match
- Query type classification
  - Identifying author queries



# Back down to earth - User Interface

- `jquery.autocomplete` Plugin
  - Behaves like an autocomplete
  - Configurable, feature rich
  - Flexible function overriding
- `jquery.trln.autosuggest` plugin
  - Home grown, layer on top of autocomplete
  - Adds some customizability
  - Usable by institutions besides just UNC.



# User Interface – Design Details

- Presentation of suggestion type
- Prescope suggestion types
- Automatic selection of search type
- Automatic deselection of search type



## Usage



**Call numbers**  
**Subject Headings**  
**LC Classification**  
**Authority Data**

...



## A search scenario...

You are at the **Library Service Center**. You want to find two titles on **History of Education** under the broader topic **"Education"**. The titles should be held by the Library Service Center and written in **Turkish**.

# Solution

The screenshot shows a library search interface with a navigation bar at the top containing tabs for 'Search', 'Advanced Search', 'Browse New Titles', and 'Browse by Call Number'. A 'Search' button is highlighted. Below the navigation bar, a search input field contains the text 'history of ed'. To the right of the input field is a dropdown menu set to 'Anywhere' and a 'Search' button. Below the input field, a list of search results is displayed, each with a subject label on the right. The results are:

- education -- history subject
- education -- united states -- history subject
- history of education in america
- history of education
- history of edgemcombe county, north carolina
- history of education in great britain
- history of education and culture in america
- history of education quarterly

Below the search results, a blue bar displays the page number '1732' and a range of numbers '[1] 2 3 4 5 6 7 8 9 10 Ne'. To the right of the search results, the text 'New Search' is visible. At the bottom right, the title 'Understanding history of education' is displayed, followed by the publication information: 'Published: Cambridge, Mass. : S' and 'Format: Book'.



## Another search scenario

You want to find three **biographies** of **American first ladies** that were published in **2010**.

# Solution1

The screenshot shows a library search interface. At the top, there are navigation tabs: "Search" (highlighted with a green speech bubble), "Advanced Search", "Browse New Titles", "Browse by Call Number", and a folder icon labeled "Ad". Below the tabs is a search bar containing the text "american first". To the right of the search bar is a dropdown menu set to "Anywhere". Below the search bar, a list of search results is displayed, each with a blue highlight on the words "american first". The results include:

- american first ladies their lives and their legacy
- american first editions bibliographic check lists of the ...
- american first editions and their prices, 1931 a ...
- the first american revolution the american colonies on ...
- american broadcasting and the first amendment

To the right of the search results, there is a "New Search" button and a pagination bar showing "1732 [1] 2 3 4 5 6 7". Below the pagination bar, there is a section titled "Understanding history" with fields for "Published:" and "Format:", and a book icon labeled "Bo".


# Solution 2

Search

Advanced Search

Browse New Titles

Browse by Call Number

 Ad

first lad|

Anywhere

first ladies

first lady

first ladies

first lady from plains

first ladies the saga of the presidents wives ...

first lady of the south

first ladys lady with the fords at the ...

first ladies of the restoration


subject

New Search

1732 [1] 2 3 4 5 6 7

Understanding history of

Published: Cambr

Format:  Bo

# Solution 3

The screenshot shows a library search interface with a navigation bar at the top containing tabs: Search (highlighted in green), Advanced Search, Browse New Titles, and Browse by Call Number. A folder icon and the text 'Ac' are visible on the right. Below the navigation bar, a search input field contains the text 'presi'. To the right of the input field is a dropdown menu set to 'Anywhere'. Below the input field, a list of search results is displayed, each with a subject label on the right. The results are:

- presidents -- united states -- biography subject
- presidents -- messages subject
- presidents -- united states -- messages subject
- presidents -- united states subject
- presidents -- united states -- election subject
- presidents spouses -- united states -- biography subject
- presidents -- united states -- history subject

On the left side of the results list, there is a vertical sidebar with a blue button labeled 'New Search' and a blue bar with the number '1732' and a list of numbers '[1] 2 3 4 5 6 7'. Below the results list, there is a section titled 'Understanding history' with the text 'Published: Camb' and 'Format: Ro'.

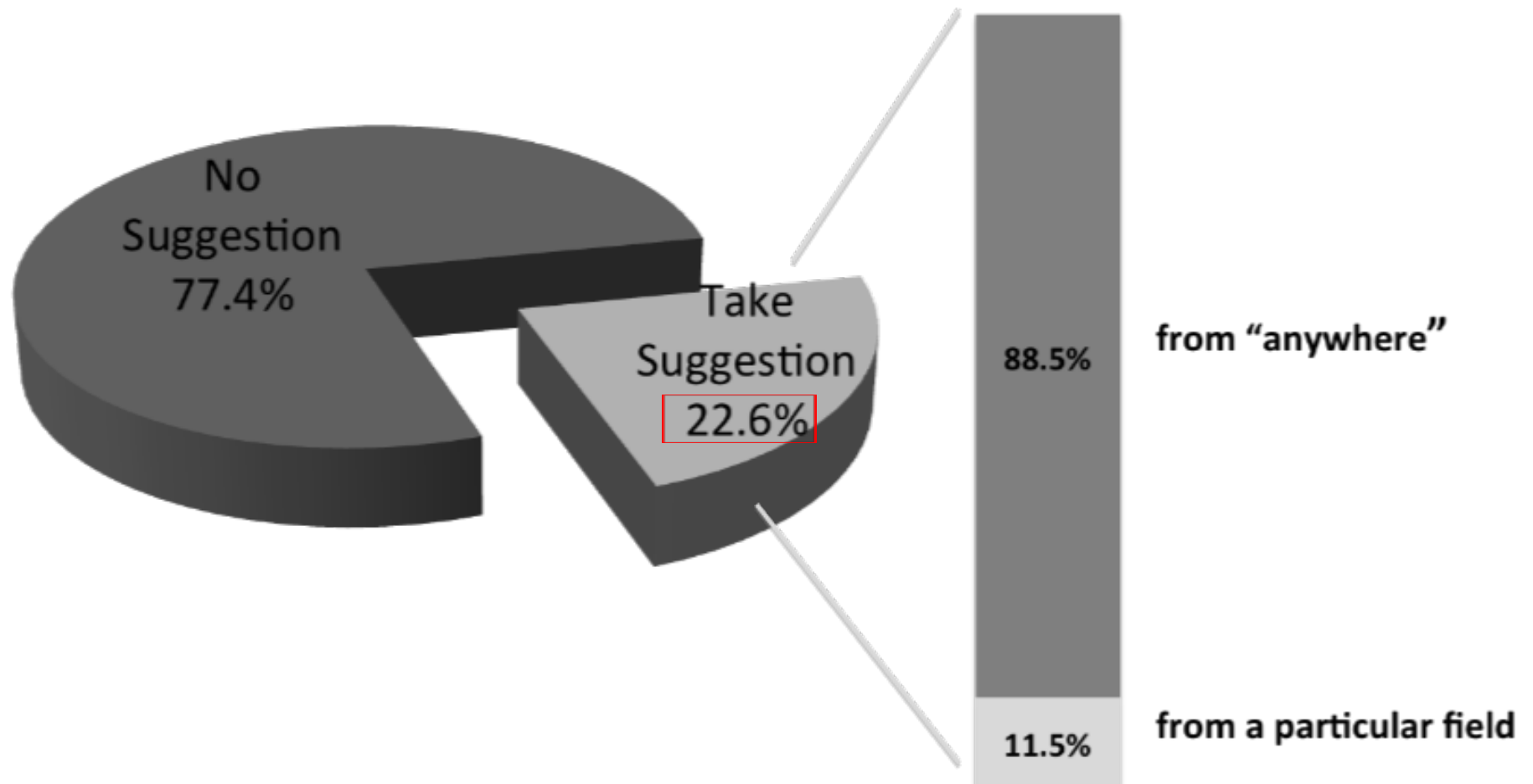
# Data

Log Dataset	Time Frame	Size	Available Fields
April 2009	30 days 4/1/2009—4/30/2009	90M raw data 378,454 useful records	IP address /data, time/URL/ reference URL/user agent
April 2010	30 days 4/1/2010—4/30/2010	109.3M raw data 412,483 useful records	IP address /data, time/URL/ reference URL/user agent

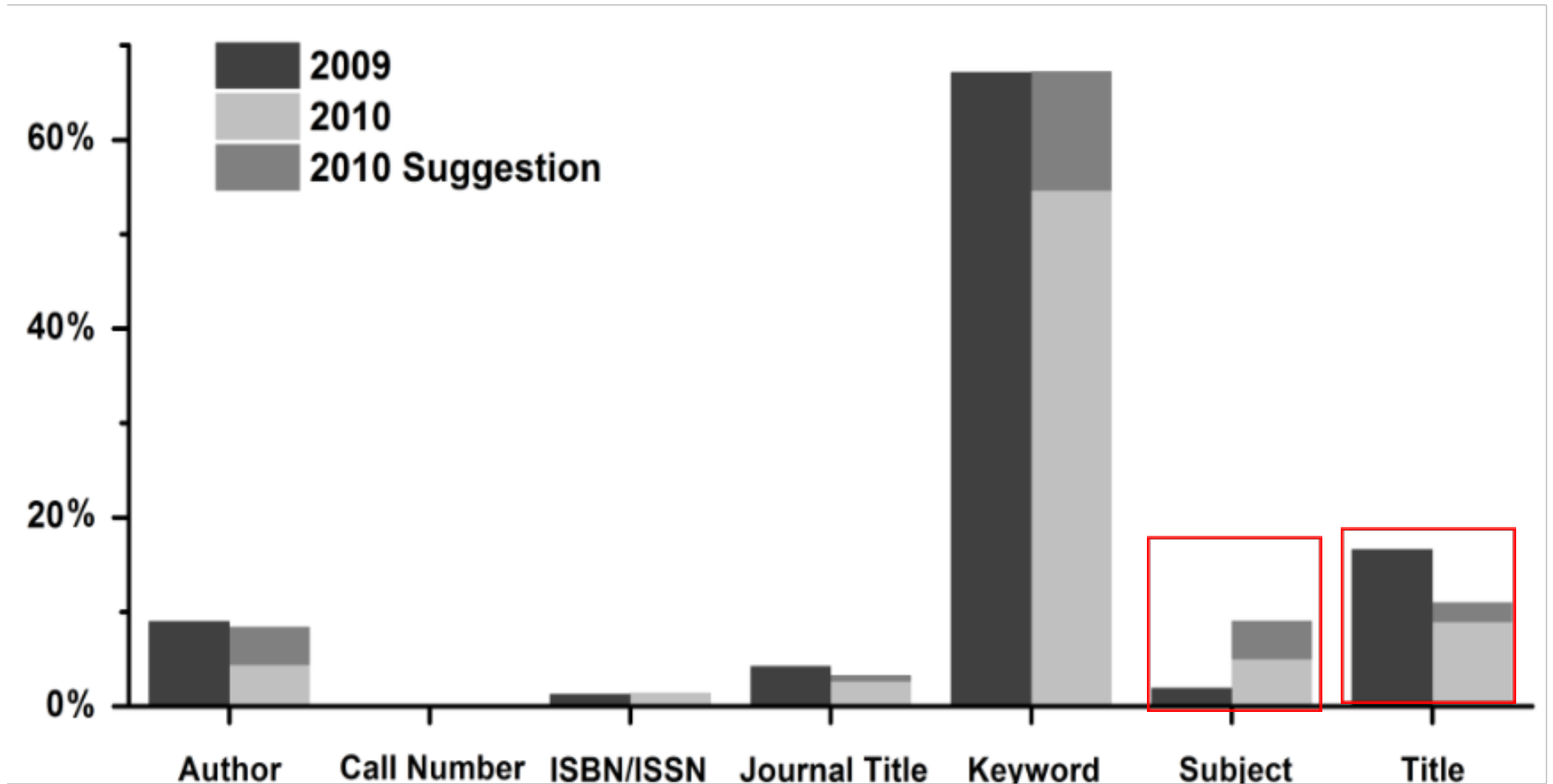
# April 2010

- 204115 queries 30 days / 6804 queries per day

# Suggestion Usage






# Changes in the search field





# Query length, # of query submissions, # of items viewed

	April 2009	April 2010
Query Length	3.1 (2.8)	3.7 (3.8) 
Number of Query Submissions per Session	4.4	2.7 
Number of Items Viewed per Session	2.3	1.6 

# Discussions

- Definition of search failure
- Fairly good uptake/No rules for using
- Boosted subject search
- “Under-specification” and “over-specification”
- Support of initial query formulation

# Conclusions

- Adoption of the feature was immediate and sustained
- Users using more authority data for searching
- Solr can handle keystroke triggered queries on 12.5 +million row text data set.
  - 84500 suggestion requests per day
- Keep design simple, clear, optional, and framed in a way users are used to.
- Fun project
- Stay tuned for Xi's user study



# Resources

- Code4lib journal article by myself and Jill Sexton:
  - <http://journal.code4lib.org/articles/3022>
- Log based usage study by Xi Niu:
  - [http://www.asis.org/asist2011/posters/165\\_FINAL\\_SUBMISSION.pdf](http://www.asis.org/asist2011/posters/165_FINAL_SUBMISSION.pdf)
- <http://docs.jquery.com/Plugins/Autocomplete>
- <http://lucene.apache.org/solr/>
- Questions?

