

Fully Automated QA System for Large Scale Search and Recommendation Engines Using Spark

Khalifeh AlJadda

www.aljadda.com

Twitter: @aljadda

Mohammed Korayem

www.cs.indiana.edu/~mkorayem/



Search Data Science



About Me



Khalifeh AlJadda

Lead Data Scientist, Search Data Science



- Joined CareerBuilder in 2013
- PhD, Computer Science – **University of Georgia (2014)**
- BSc, MSc, Computer Science, **Jordan University of Science and Technology**

Activities:

- Founder and Chairman of CB Data Science Council
- Frequent public speaker in the field of data science
- Creator of [GELATO](#) (Glycomic Elucidation and Annotation Tool)

CAREERBUILDER™ Search by the Numbers



100 million +
Searches per day



1,5 billion +
Documents indexed and
searchable



500+
Search Servers



30+
Software Developers, Data
Scientists + Analysts



1
Global Search
Technology platform

Powering 50+ Search Experiences Including:



Search Pro
*(Search-CareerBuilder RDB,
Recruitment Edge, Supply & Demand)*



Talentstream Supply & Demand
(Supply & Demand Portal)



CAREERBUILDER™

MiracleWorkers.com
heartbeat



Small Business Resume Database
(Search Basic, RDB Basic)



Candidate Sourcing Platform

workinretail.com
Retail Jobs. Retail talent.

WorkInNursingJobs.com

sologig.com



Talentstream Engage
(Talent Network)



Broadbean Resume Search
(Multi-vendor Resume Search)

StaffNurse.com
*Nursing & Healthcare Jobs
A CareerBuilder Company*

phonemplo.com

HEADHUNTER.com



Talentstream Recruit
(CareerBuilder1)



Talentstream Gather
(Talent Gather)

Lesjeuds.com

RecruLex.com
scal

MoneyJobs.com

OILANDGAS
JOBSEARCH

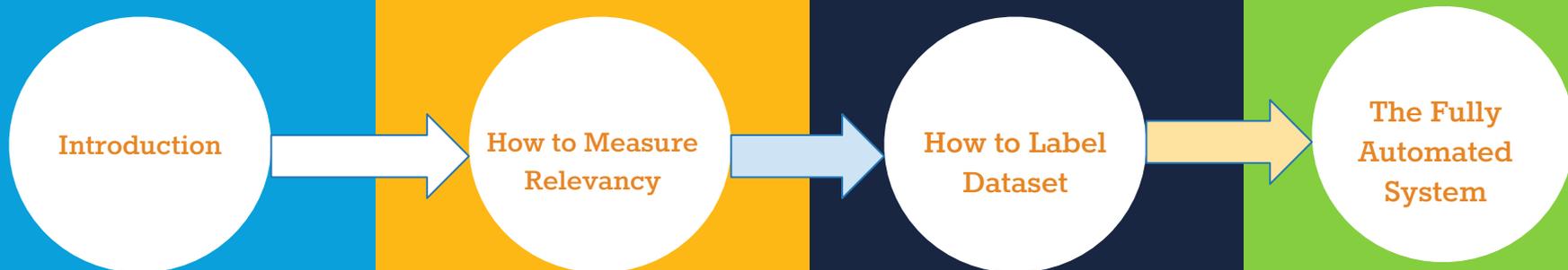
e-recrut.com
Emploi Commerce - Vente

CAO
emplois.com

WorkInTherapyJobs.com

...and many more

Talk Flow





Introduction



What is Information Retrieval (IR)?

Information retrieval (IR) is **finding** material (usually documents) of an **unstructured** nature (usually text) that **satisfies** an information **need** from within **large collections** (usually stored on computers).*

*introduction to information retrieval: <http://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf>

Information Retrieval (IR) vs Relational Database (RDB)

	RDB	IR
Objects	Records	Unstructured Documents
Model	Relational	Vector Space
Main Data Structure	Table	Inverted Index
Queries	SQL	Free text

The inverted index

What RDB would store:

Document ID	Content Field
doc1	once upon a time, in a land far, far away
doc2	the cow jumped over the moon.
doc3	the quick brown fox jumped over the lazy dog.
doc4	the cat in the hat
doc5	The brown cow said "moo" once.
...	...

How the content is INDEXED into Inverted Index:

Term	Documents
a	doc1 _[2x]
brown	doc3 _[1x] , doc5 _[1x]
cat	doc4 _[1x]
cow	doc2 _[1x] , doc5 _[1x]
...	...
once	doc1 _[1x] , doc5 _[1x]
over	doc2 _[1x] , doc3 _[1x]
the	doc2 _[2x] , doc3 _[2x] , doc4 _[2x] , doc5 _[1x]
...	...

Vocabulary

Relevancy: Information need satisfaction

Precision: Accuracy

Recall: Coverage

Search: Find documents that match a user's query

Recommendation: Leveraging context to automatically suggest relevant results



How to Measure Relevancy



Motivation

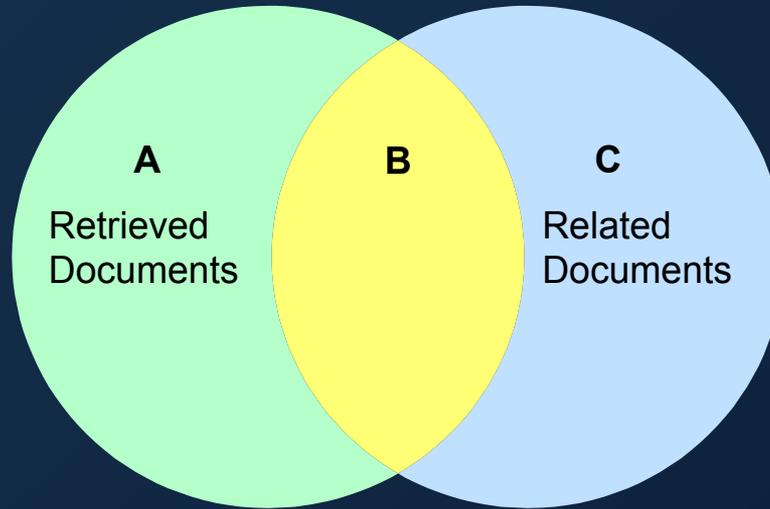
Users will turn away if they get irrelevant results

New algorithms and features need test

A/B test is expensive since it has impact on the end users

A/B test requires days before a conclusion can be made

How to Measure Relevancy?



Precision = B/A

Recall = B/C



$$F1 = 2 * (Prec * Rec) / (Prec+Rec)$$

Assumption:

We have only 3 jobs for aquatic director in our Solr index

$$\text{Precision} = 2/4 = 0.5$$

$$\text{Recall} = 2/3 = 0.66$$

$$\text{F1} = 2 * (0.5 * 0.66) / (0.5 + 0.66) = 0.56$$

Problem:

Assume **Prec = 90%** and **Rec = 100%** but assume the 10% irrelevant documents were ranked at the top of the results

is that OK? 

Aquatic Director Jobs

[Create Job Alert](#)

Keywords: Location: Posted within:

Filter By

7 Jobs Found

Sort by: [Job Title](#) | [Location](#) | [Relevance](#)

Project Director - Manager, Construction-Engineering

Job type: Full-Time
Project Manager - Construction Swimming Pools, Water Parks and Aquatic Facilities
Natare is seeking an individual for a project management position...
[Save Job](#) [Email Job](#)



AQUATICS DIRECTOR

Job type: Full-Time
The Lakota Family YMCA is seeking a team-oriented, motivated professional for our full time Aquatics Director position. The successful candidate wi...
[Save Job](#) [Email Job](#)



Aquatics Director

Job type: Full-Time
YMCA of Walla Walla is seeking a team-oriented, motivated professional for our full time Aquatics Director position. The successful candidate will...
[Save Job](#) [Email Job](#)



Director, Human Resources

Job type: Full-Time | Pay: \$130k - \$170k/year
Moore & Associates has been retained by Motion Picture & Television Fund in Woodland Hills, CA to conduct the following recruitment: Director, Huma...
[Save Job](#) [Email Job](#)



Discount Cumulative Gain (DCG)



Rank	Relevancy
1	0.95
2	0.65
3	0.80
4	0.85

Rank	Relevancy
1	0.95
2	0.85
3	0.80
4	0.65

$$DCG_k = \sum_{i=1}^k \frac{2^{rel_i} - 1}{\log_2(i + 1)}$$

$$NDCG_k = \frac{DCG_k}{IDCG_k}$$

- **Position** is considered in quantifying relevancy.
- **Labeled dataset** is required.



How to Label Data Set?



How to get labeled data?

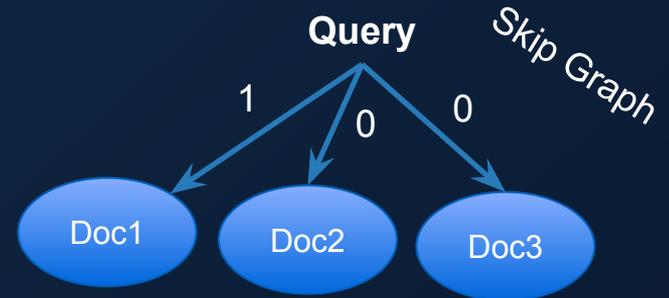
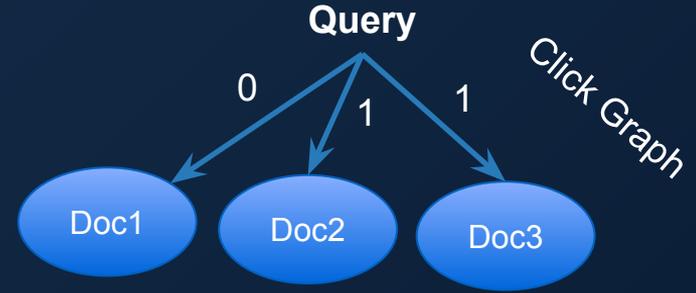
- Manually
 - Pros:
 - Accuracy
 - Cons:
 - Not scalable
 - Expensive
 - How:
 - Hire employees, contractors, or interns
 - Crowd-sourcing
 - Less cost
 - Less accuracy
- Infer relevancy utilizing implicit user feedback

How to infer relevancy?

Query



Rank	Document ID	
1	Doc1	X
2	Doc2	✓
3	Doc3	✓
4	Doc4	?



$$C_{Doc_i} = \sum_{Doc_i} \text{click}$$

$$S_{Doc_i} = \sum_{Doc_i} \text{skip}$$

$$rel_{Doc_i} = \frac{C_{Doc_i}}{S_{Doc_i} + C_{Doc_i}}$$

Query Log

Field	Example
Query ID	Q1234567890
browser ID	B12345ABCD789
Session ID	S123456ABCD7890
Raw Query	Spark or hadoop and Scala or java
Host Site	US
Language	EN
Ranked Results	D1, D2, D3, D4, .. , Dn

Action Log

Field	Example
Query ID	Q1234567890
Action Type*	Click
Document ID	D1
Document Location	1

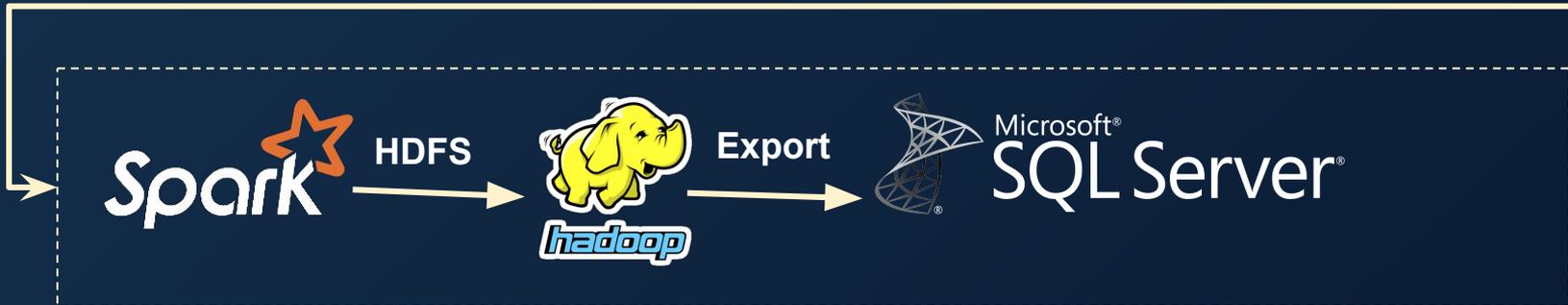
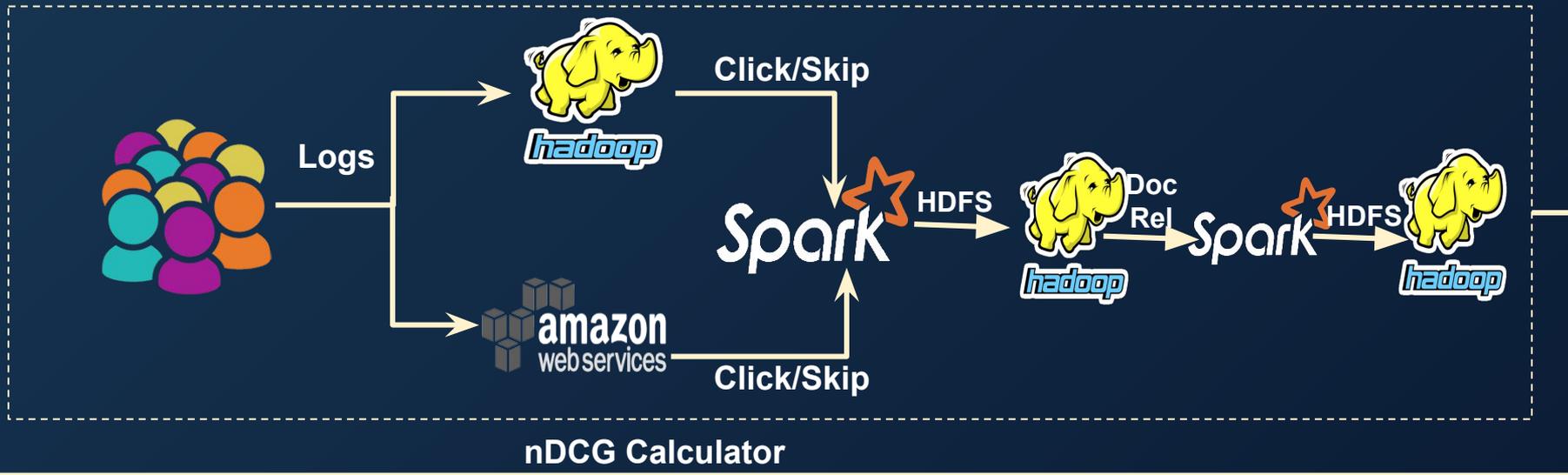
*Possible Action Types: Click, Download, Print, Block, Unblock, Save, Apply, Dwell time, Post-click path



The Fully Automated System



System Architecture



ETL

Field	Example
Query ID	Q1234567890
browser ID	B12345ABCD789
Session ID	S123456ABCD7890
Raw Query	Spark or hadoop and Scala or java
Ranked Results	D1, D2, D3, D4, .. , Dn



Field	Example
Query ID	Q1234567890
Action Type*	Click
Document ID	D1
Document Location	1

Keyword	DocumentID	Rank	Clicks	Skips	Popularity
---------	------------	------	--------	-------	------------

Keyword	DocumentID	Relevancy
---------	------------	-----------

Noise Challenge

At least 10 distinct users need to take an action on a document to consider it in the nDCG calculation.

Any skip followed clicks on different sessions from the same browser ID is ignored.

Actions beyond Clicks weight more than Clicks. For example, we count Download as 20 clicks, and Print as 100 clicks

Accuracy

500 resumes had been manually reviewed by our data analyst. The accuracy of the relevancy scores calculated by our system is

96%

Dataset by the Numbers



19 million +
Search logs



250,000+
Tagged resumes



100,000+
Distinct Queries



10+
Distinct users per query

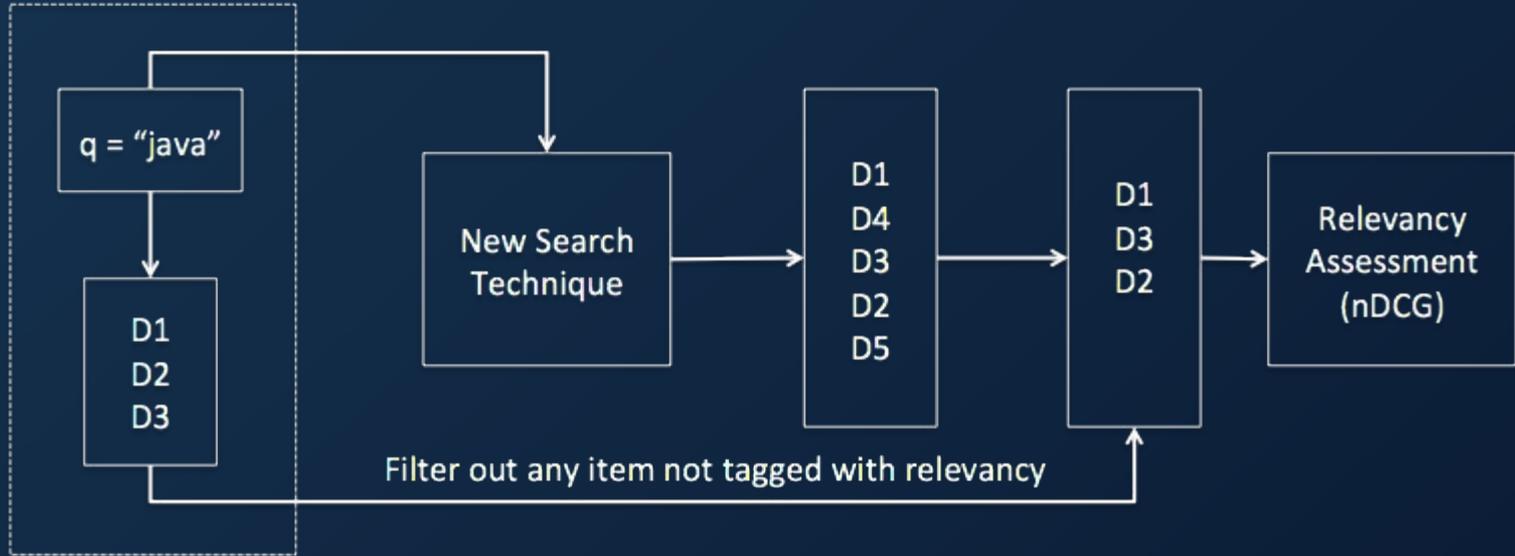


7
Times a week

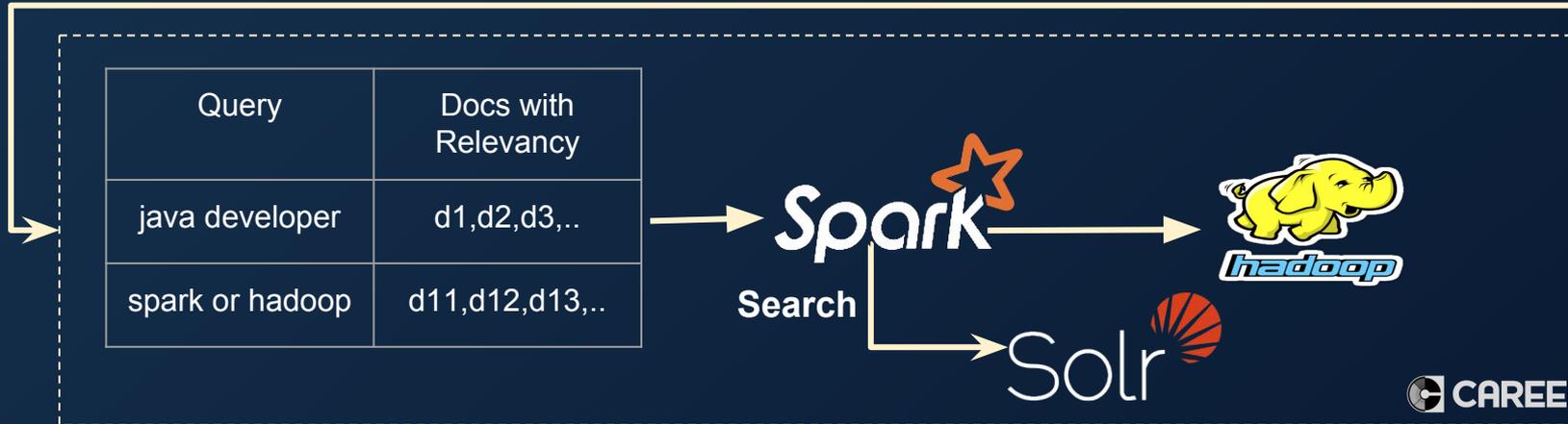
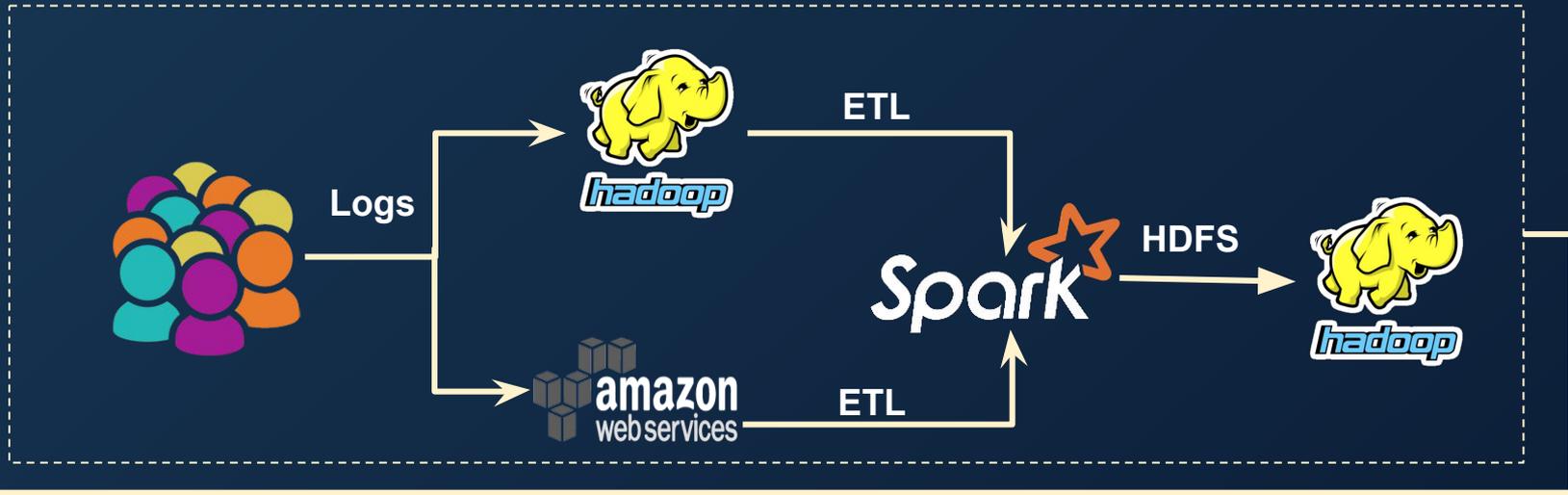


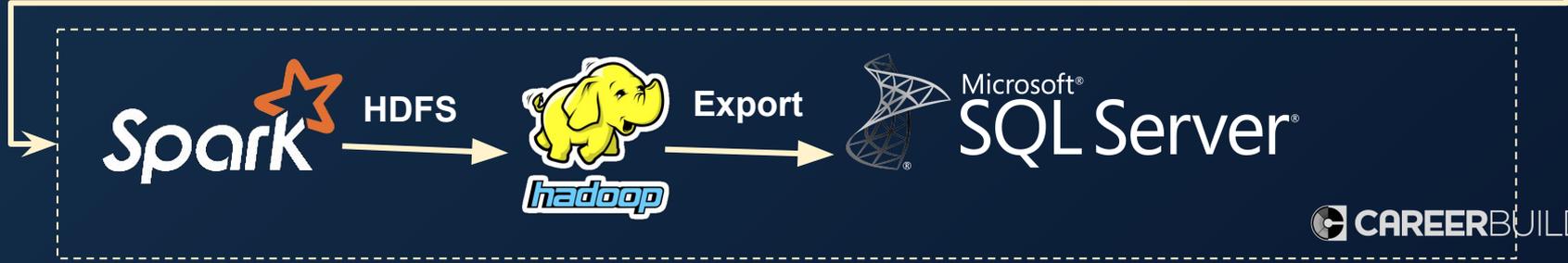
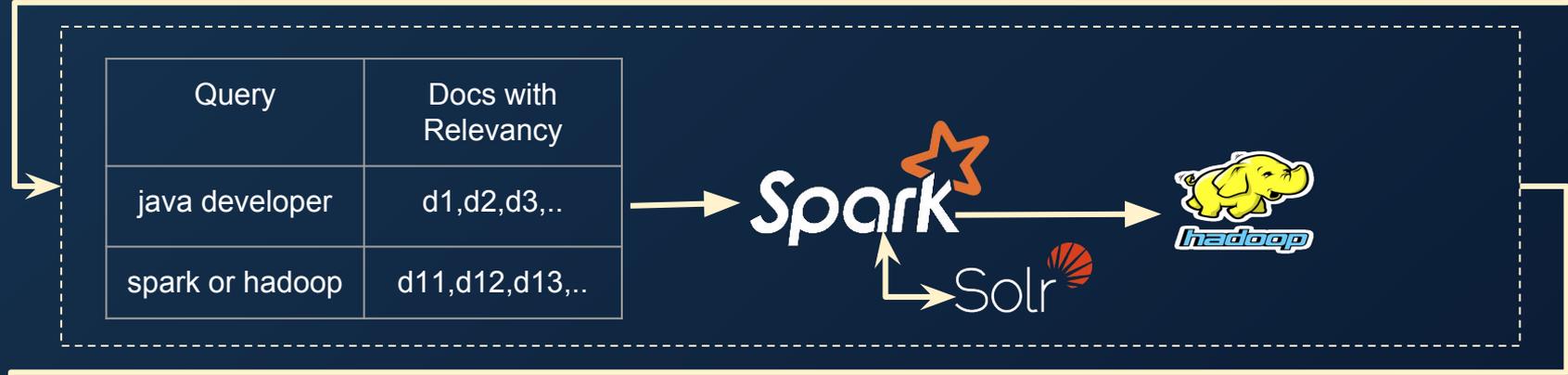
Microsoft®
SQL Server®

Query Synthesizer

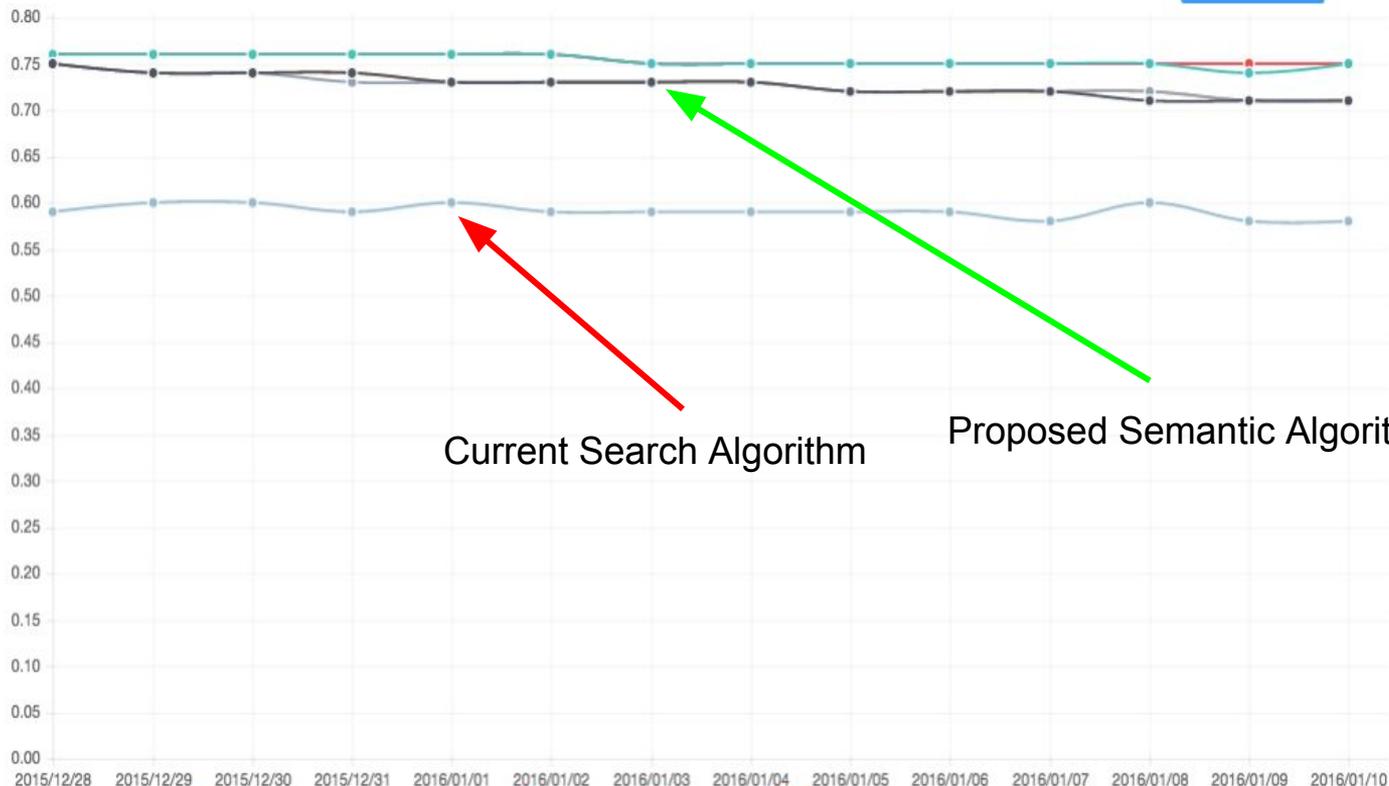


Synthesize Queries





DCG TRENDS 7

7 checked -

Current Search Algorithm

Proposed Semantic Algorithms



© PA

Quick Win



Relevancy Guard

Top Ranked, Low Relevancy

These jobs are top ranked in these searches, but with a low relevancy score. What would you like to do with them?

Physical Education Teacher

Rank: 1 | Relevancy Score: 0.0

Search query: [Education Jobs In Rochester,Ny](#)

 Delete

 Bump down

 Keep as is

Home Health Instructor – Chinese / Mandarin Speaking RN

Rank: 1 | Relevancy Score: 0.0

Search query: [Chinese Jobs](#)

 Delete

 Bump down

 Keep as is

SERVER / RESTAURANT / HOSPITALITY EXPERIENCE - CUSTOMER RELATIONS REPS

Rank: 1 | Relevancy Score: 0.0

Search query: [Hospitality Jobs In Atlanta,Ga](#)

 Delete

 Bump down

 Keep as is

Licensed Practical Nurse - LPN - Various Specialties Needed!

Rank: 1 | Relevancy Score: 0.0

Search query: [Nursing Jobs In Macon,Ga](#)

 Delete

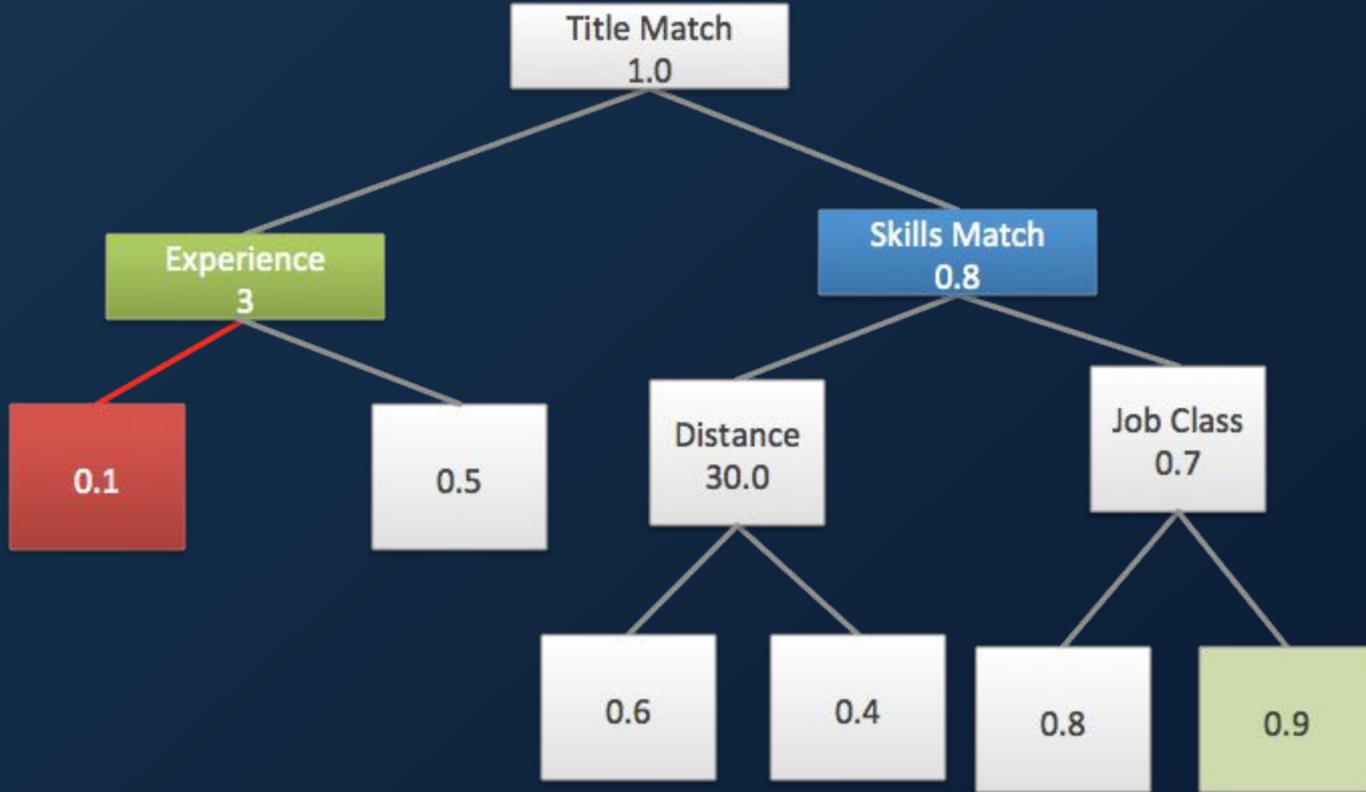
 Bump down

 Keep as is

Learning to Rank (LTR)

- It applies machine learning techniques to discover the best combination of features that provide best ranking.
- It requires labeled set of documents with relevancy scores for given set of queries
- Features used for ranking are usually more computationally expensive than the ones used for matching
- It works on subset of the matched documents (e.g. top 100)

LambdaMart Example

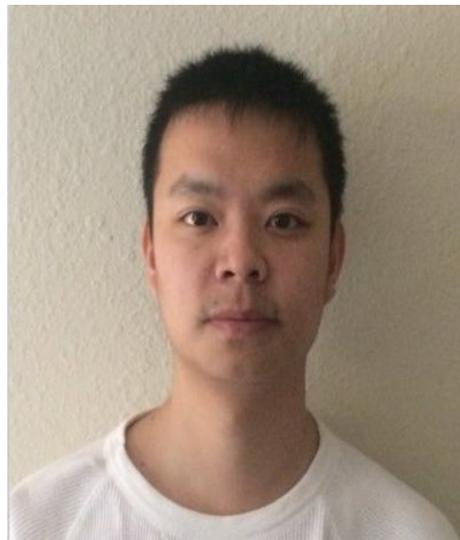




Mohammed Korayem



Hai Liu



Chengwei Li



David Lin

Credit





Thank You!

Khalifeh AlJadda

www.aljadda.com

Twitter: @aljadda

Search Data Science

