Limited Access Document

Search Engine Profiles

Gilbane Seminar

December 2, 2004

ARNOLD
Information
Technology

Discuss 11 Engines and Google

- “Safe” choices
- Hot engines
- Special purpose engines
- Google Appliance

New Book on Enterprise Search

http://www.cmswatch.com/EntSearch/

Safe Choices

Name	*Description*	*Strength*	*Disadvantage*	*License Fee*
Autonomy	Minimal human intervention. Uses Bayesian math.	When properly set up, licensees like its discovery features	Diverse content can affect precision	$300,000/year
Convera	Text, database, and image retrieval. Includes taxonomies for finance,	Can provide Google simplicity or Boolean complexity	Financial condition of the company	Begins at $50,000
FAST Search	ESP handles structured, unstructured, and Web data. Clustering available	With appropriate computing resources, among the fastest engines	Original coding required for certain functions	$250,000/year
Verity	Ultraseek for basic indexing. K2 for text, database and structured	High profile, accepted brand	Verity consulting needed to deploy engine	$35,000 excluding services

Hot Engines

Name	*Description*	*Strength*	*Disadvantage*	*License Fee*
Endeca	Generates metadata to support clustering and Google-style queries	Can be integrated into work flows	Endeca professional services needed	$200,000
InQuira	NLP, taxonomies, and learning from user behavior	Tailored for customer service applications	Human tuning required to eliminate incorrect answers	250000
iPhrase	XML repository plus taxonomy and key word indexing	Documents generated from the repository	Storage and computational infrastructure	$250,000
TripleHop	Generates metadata to support clustering and Google-style queries	Effective clustering and metadata generation	Computational infrastructure	$250,000

Special Purpose Engines

Name	*Description*	*Strength*	*Disadvantage*	*License Fee*
Blossom	Managed service or on-site install	Fast, easy-to-configure	Limited file type support	$12,000
Mondosoft	Search plus usage tracking functions	Plugs into Microsoft servers	Performance in some implementations	$35,000
Vivisimo	Metasearch engine	Can deduplicate and cluster hits	Computational infrastructure	$250,000

Federated Search

- Pioneered by Vivisimo
- Federated Search
  - "Meta Search" engine
  - Queries other indexes and clusters common results
- A "feature" in leading search engines
  - Verity, FAST
- Many low cost, practical uses

Federated Engines

- A9.com
- Clusty
- Virgilio
- Lycos
- FirstGov moving in this direction

Google Appliance

- Plug and play search
  - Minimal configuration
  - Expansion easy… “Buy” more Appliances
- Pricing begins at $32,000
  - Each pizza box can index 150,000 documents
  - More than 150,000, more Appliances
- User can define collections

Google Appliance Benefits

- Easy to set up
- Can do a Web site search or an enterprise search
- Minimal system administration
- Users “love” Google
- Scaling painless
- Indexing speed slower than delivering results of the query
- Customized version of Linux

Google Appliance Drawbacks

- Minimal configuration
- Pricing thresholds not well understood
- An enterprise class system that can handle one million documents is about $250,000
  - Comparable to Autonomy, Convera, FAST, and Verity
  - Precision and recall acceptable but not like the Web version

General: The Services Issue

- Most search engines are “kits” or “parts”
- Assembling, setting up, tuning, and upgrading require insider knowledge
- Verity derives more than ½ its revenue from services
- Autonomy relies on resellers
- Other vendors have different policies
  - Search engine pays an engineer $60K
  - Customer pays $240,000 with mark up of 4X

The Performance Issue

- **Hard disc space, including scratch space during indexing and index updates**
- **RAM—as much as possible in each machine in the search system**
- **Processors. Two schools**
  - **Google approach. Commodity machines**
  - **IBM approach. Carrier class server machines**
  - **Computational power needed in either approach**