1

image ® Rolls-Royce

Hybrid Search: Effectively Combining Keywords and Semantic Search

Ravish Bhagdev, Sam Chapman, Fabio Ciravegna, Vitaveska Lanfranchi
web Intelligence technology lab,
nlp group, department of computer science,
university of sheffield

Daniela Petrelli
department of Information studies,
university of sheffield

Outline of Talk

- Aim of paper
  - Search for what and in what conditions?
- Hybrid Search as a way to overcome limitation of classic semantic search
- Implementing Hybrid Search into K-Se
- Experimental Evaluation in vitro and ir
- Conclusion and future work

3 mins+3 slides

7min+9slides

4 mins+5slides

8 mins+15 slides

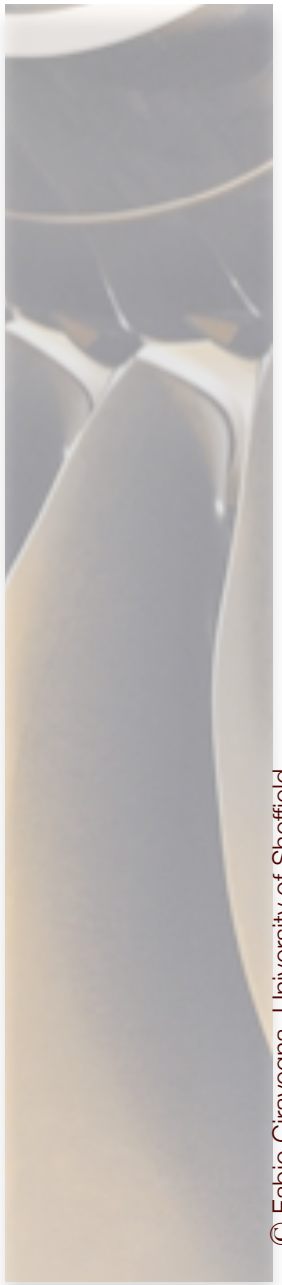
3 mins+2 slides

Aim of the paper

- We propose a search method
  - Designed for the Semantic Web
    - Seen as a collection of both documents and metadata,
  - Designed to achieve two tasks:
    - Document retrieval: searching for documents using concepts or keywords of interest
    - Knowledge retrieval: retrieving facts from a knowledge base (i.e. triples)

Metadata and User Needs

- Differently from [1, 2, 3, 4, 9],
  - We expect metadata to cover only partially the user information needs
    - Reasons:
      - limitations in the ontology wrt user needs
      - limitations in the annotation capabilities
        - i.e. limitations in IE capabilities
      - metadata unavailable for a specific document

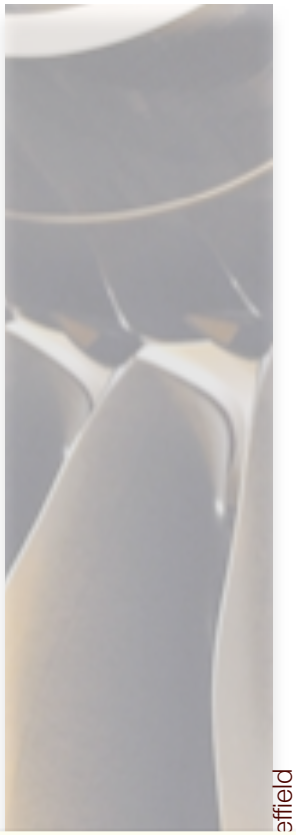


Pure Semantic Search (OS)

- Semantic search as metadata-based search defined according to an ontology,
  - Annotations are unambiguous
    - OS Does not suffer from ambiguity and synonym issues of keyword-based systems (KS)
- Issue:
  - OS can fail to encompass user information needs
    - When metadata does not completely cover user needs

Hybrid Search

- We propose a model of searching combining
  - the flexibility of keyword-based retrieval
  - querying and reasoning capabilities of semantic search
- HS is formally defined as:
  - the application of semantic (metadata-based) search for the parts of the user queries
    - where metadata is available
  - the application of keyword-based search for the parts not covered by metadata.



- But also it must leave freedom to users to chose among the two paradigms!
  - As we will see users make a creative use of it

Queries in Hybrid Search

differently from other approaches (e.g. [9]), in HS conditions on metadata and keywords coexist.

- Any boolean combination of three t
  conditions
  - pure semantic:
    - via unique identification of objects/relations
      - e.g. via URIs or unique identifiers
  - keyword-based
    - matching on the whole document
  - keyword-in-context
    - matching keywords only within portion of documents semantically annotated with a specific type or instance

Queries in Hybrid Search

differently from other approaches (e.g. [9]), in HS conditions on metadata and keywords coexist.

- Any boolean combination of three t
  conditions
  - pure semantic:
    - via unique identification of objects/relations
      - e.g. via URIs or unique identifiers
  - keyword-based
    - matching on the whole document
  - keyword-in-context
    - e.g. it enables searching for the string "fuel" but only in the context of all the text portions annotated with the concept affected-engine-part [14]

Example of Hybrid Query

∀x,y,z /

(discoloration y) & (located-on y x) & (component x)

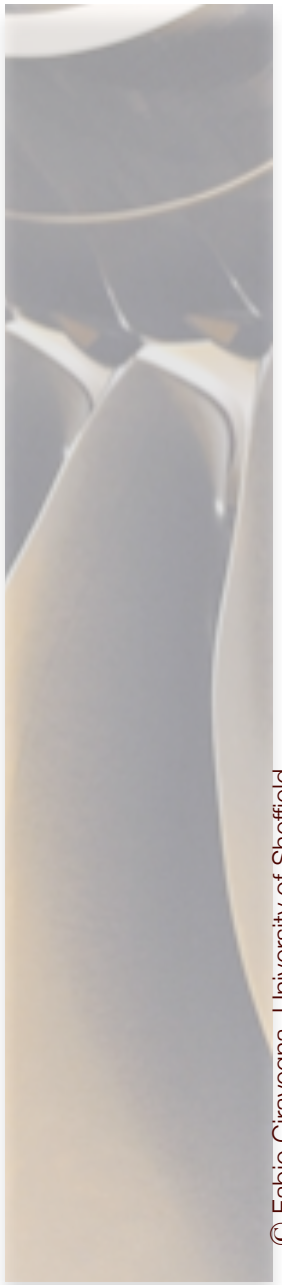
Querying Metadata

& (provenance-text-contains x “blade”)

Keyword in Context Query

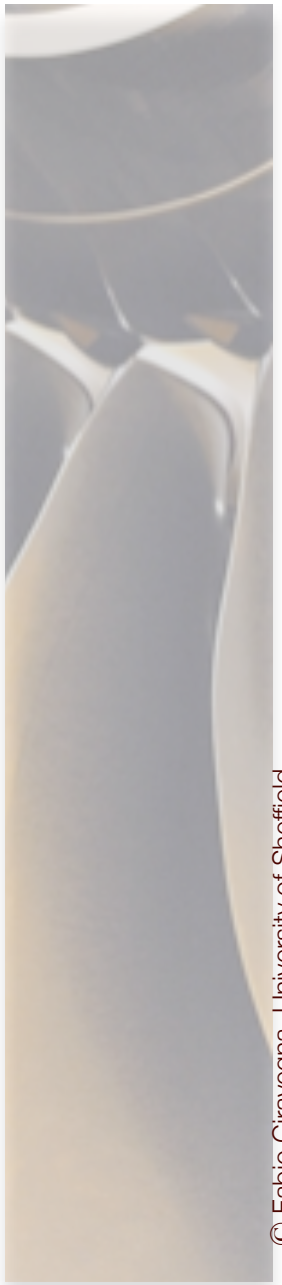
& (contains z “trailing edge”) & (document z) & (provenance x z)

Keyword-based Query



Implementing HS: Indexing

- Documents are indexed using a standard keyword–based engine such as SolR
- Facts (e.g. extracted by an IE system) are stored in a Knowledge Base
  - e.g. a triple store like Sesame2 in the form of RDF triples.
- Provenance of facts recorded
  - E.g. As triples connecting
    - the facts' URIs and those of the document of origin
    - the facts' URIs and the original strings used in the documents



Implementing HS: Retrieval

- Query is parsed and the different components (keywords, keywords-in-context and metadata) identified
  - keyword matches ➜ traditional information retrieval system
  - metadata searches
    - Translated into a query language like SPARQL
    - Sent to a triple store
  - keywords-in-context queries
    - matched with provenance of annotations in documents
      - E.g. Using SPARQL and a triple store
- Finally, results are merged, ranked and displayed

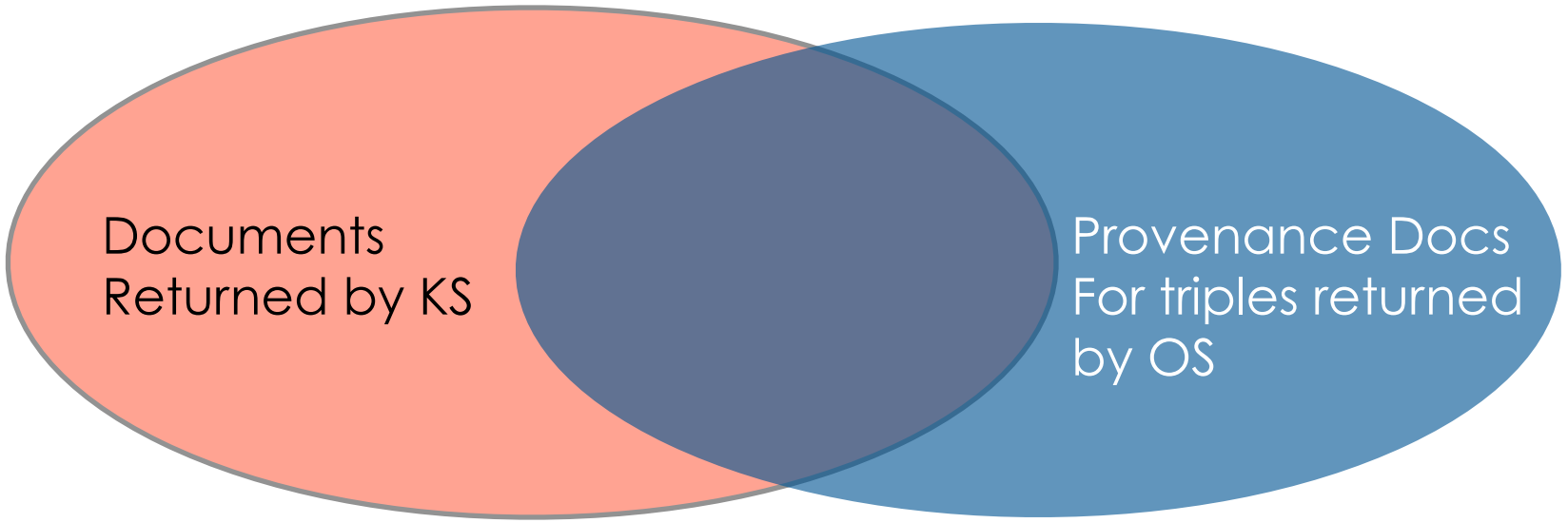
Result Merging

- Merging keyword and semantic results is not straightforward
  - Keyword matching returns an ordered set of URIs of documents
  - a semantic search returns an unordered set of assertions < subj, rel, obj>
- Merging is a different task if:
  - Document Searching
    - Returns documents
  - Knowledge Searching
    - Returns triples

Merging results

- Provenance of triples returns document ids for triples (URIs)
  - Document Searching:
    - Provenance URI set is intersected with URIs of documents returned by keywords
    - `HybridSearchUriSet= KSDocUriSet ∩ OSDocUriSet`

I won't mention ranking here

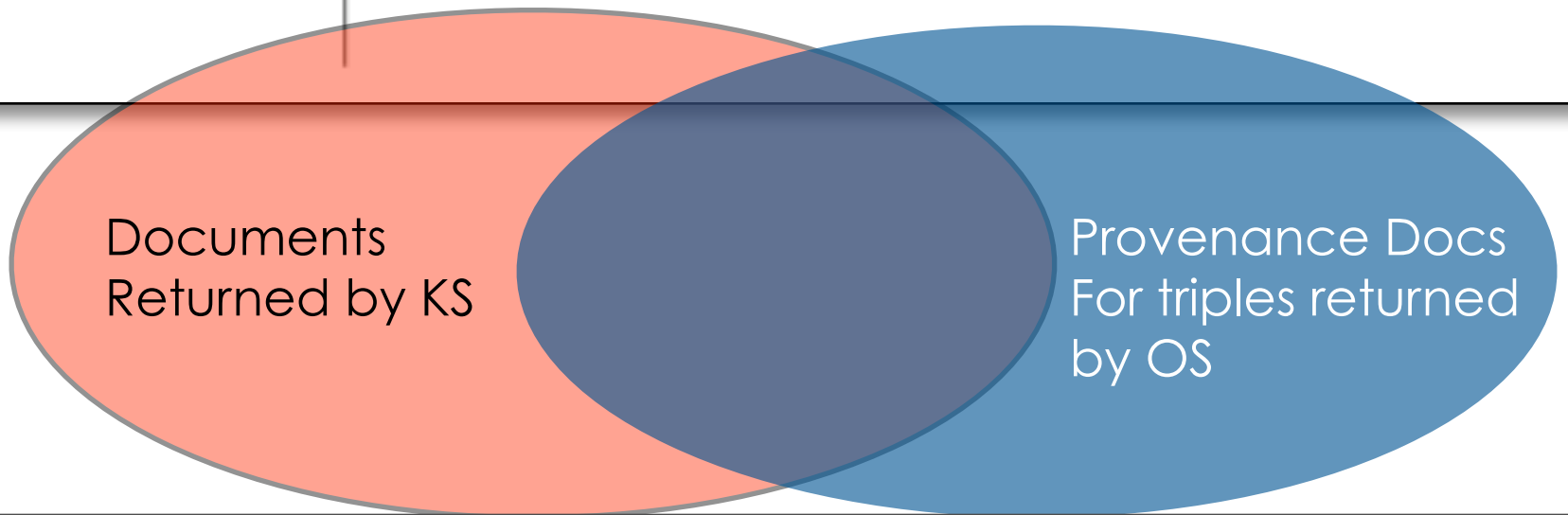


Merging results

- Provenance of triples returns document ids for triples (URIs)
  - Knowledge Searching
    - Triples returned by semantic search are filtered t
      remove those whose provenance does not point to
      any of the documents returned by the keywords

I won't mention ranking here

$$HSTripleSet = \{ \text{All } triples \in OSTripleSet \text{ Where } Provenance(triple^i) \in KSDocUriSet \}$$



Expected effect of HS: Document Searching

- With respect to OS
  - Recall expected to increase
    - Use of keywords where metadata is missing enables to answer otherwise impossible queries
  - Precision may suffer because of polysemy
- With respect to KS
  - Precision and recall expected to increase
    - Ambiguity and synonymity are dealt with by semantic search when available
      - Higher recall and precision
    - As keywords are combined with metadata in the same query, the context given by the available metadata helps in disambiguating keywords as well
      - higher precision

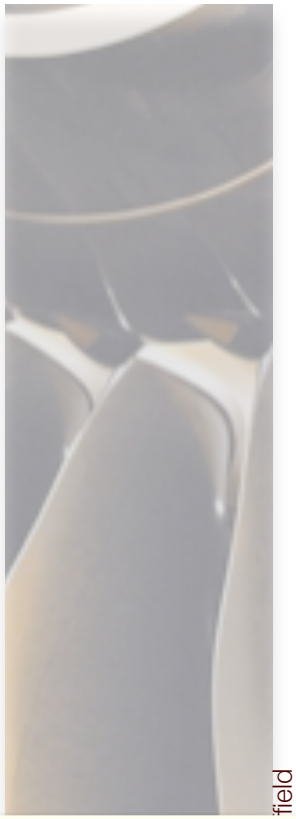
Expected effect of HS : Knowledge Searching

- With respect to OS
  - Precision increased
    - Use of keywords where metadata is missing enables more precise queries
      - although less precise than the ideal ones

OR

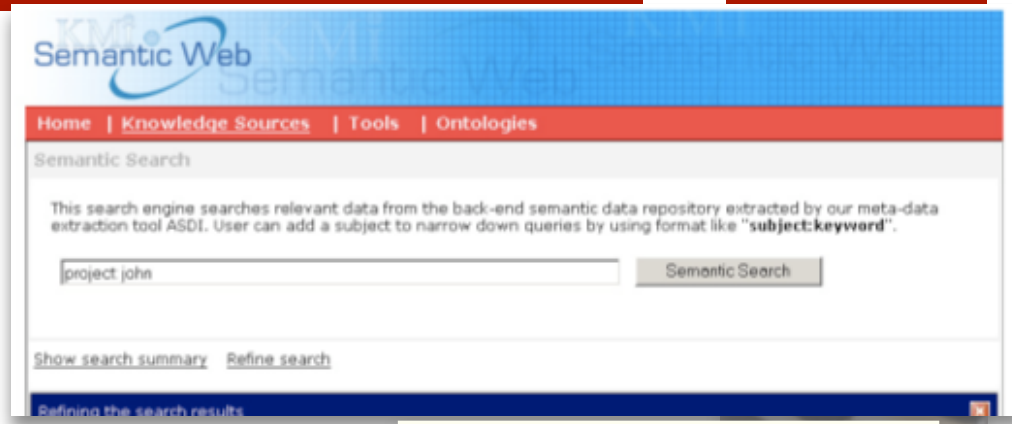
  - Recall increased
    - Use of keywords where metadata is missing enables to answer otherwise impossible queries
  - Precision may suffer because of polysemy

- With respect to KS
  - KS does not cover Knowledge Searching



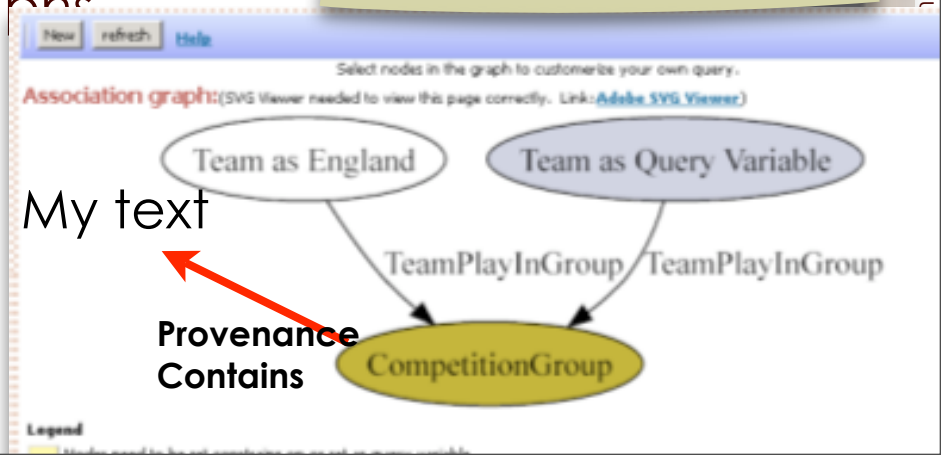
Next slide:
We have implemented a version to confirm our expectation

Implementing HS: What Search Strategy?



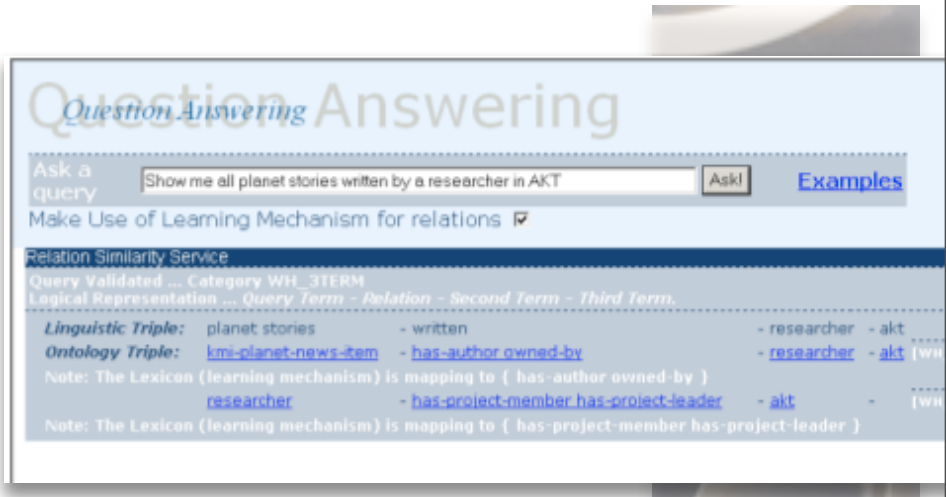
- Keyword-based approaches
  - Require translating all the keywords in order to perform the query
    - E.g. SemSearch
    - HS implemented by replacing keywords in the query with c the ontology when possible while leaving the rest for pure k based searching
    - Keywords in context rather difficult
- View-based approaches
  - Based on querying by building visual graphs
    - E.g. Falcon
    - HS support by adding two arc types
      - document-contains
      - Object description contains

Go through this and next slide very quickly !!

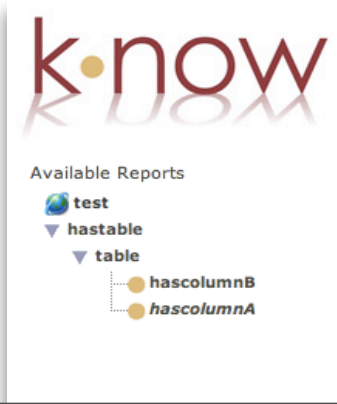


Search Strategy (ctd)

- A natural language approach
  - E.g. Aqua
  - HS suported by recognising expressions like
    - "and the document contains…"
    - And its description contains

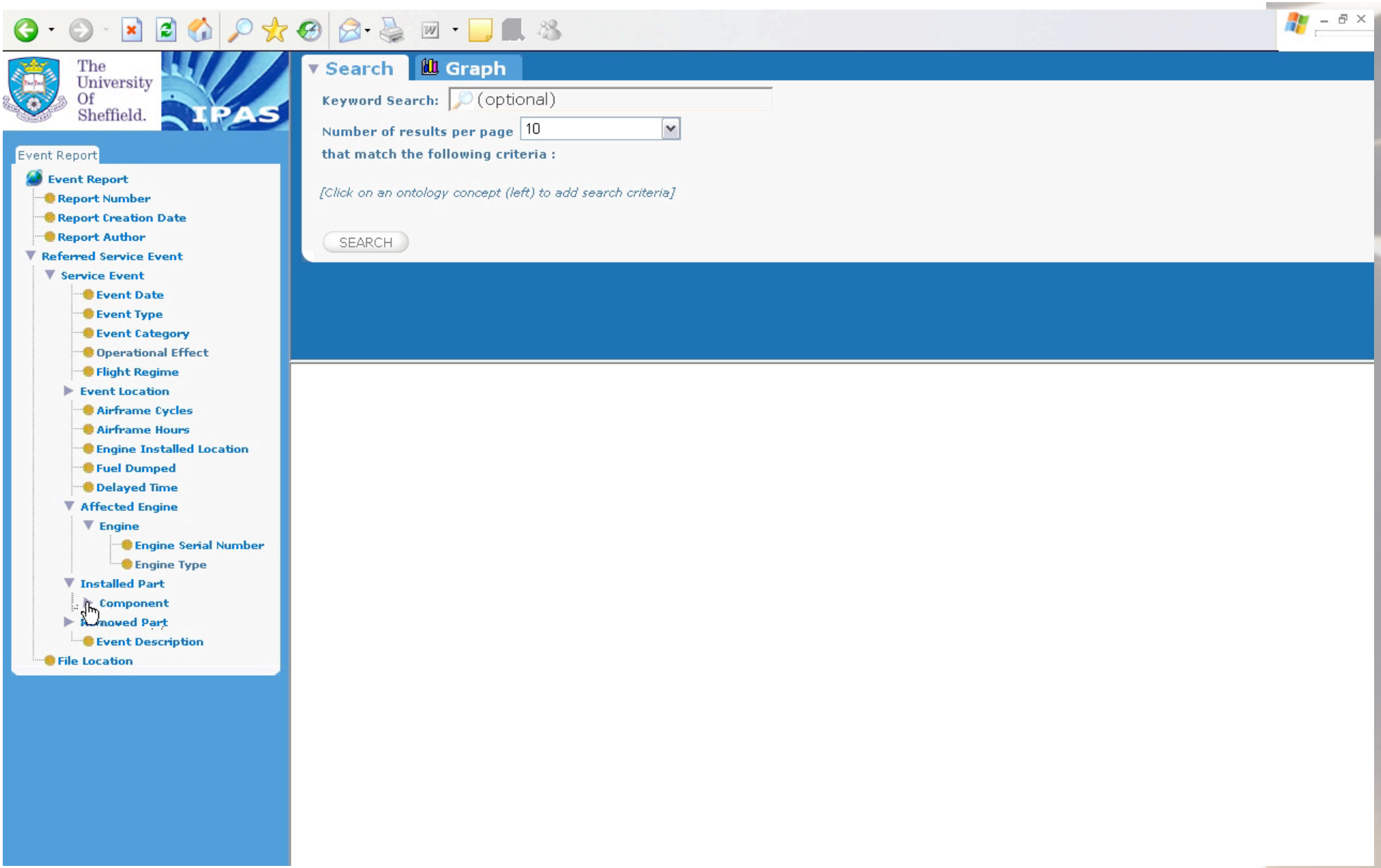


- Form-based approaches
  - HS supported by introducing
    - Keyword Search field
    - Enable keyword Matching on fields

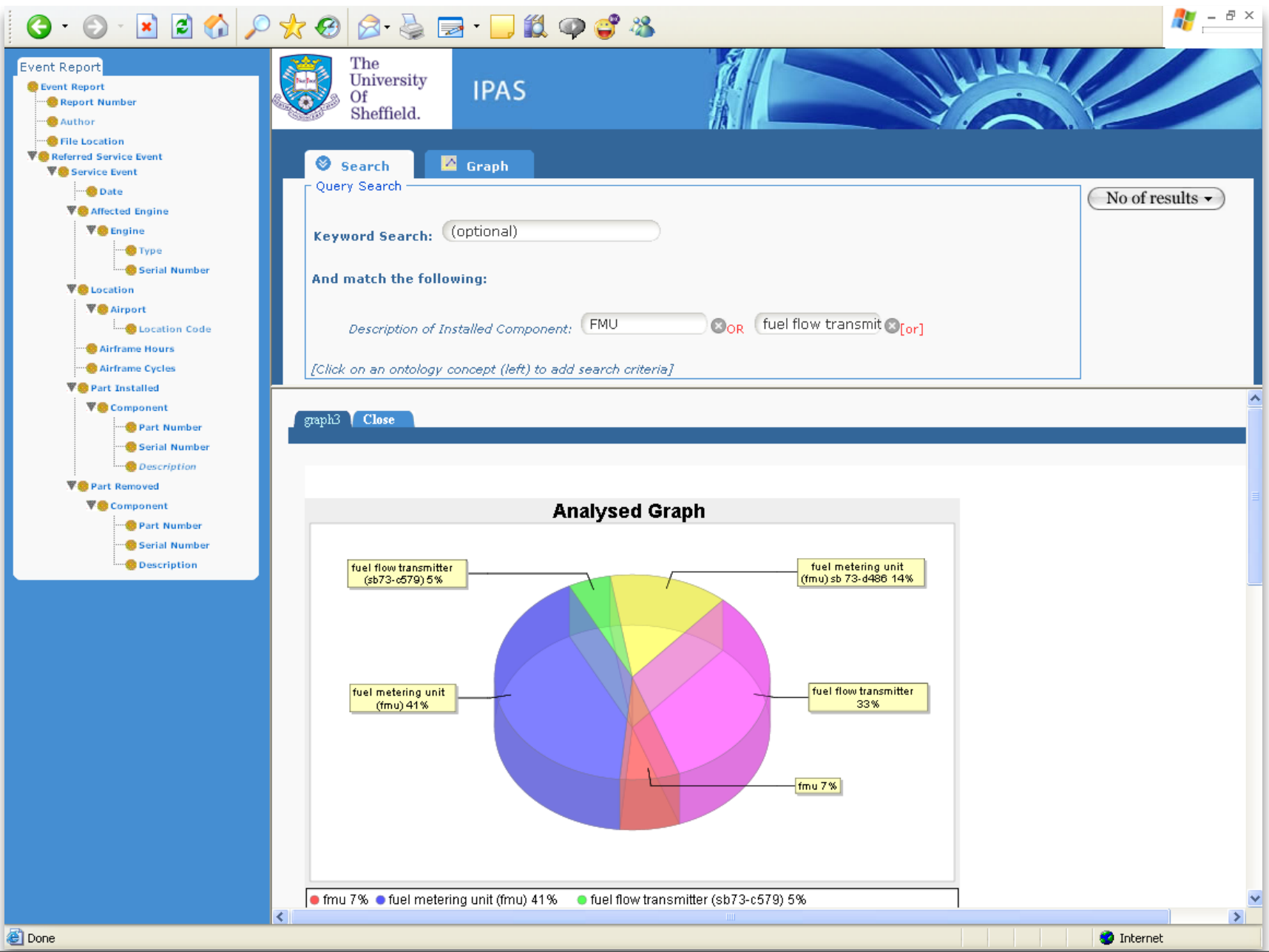


- Form-based implementation of hybrid search initially created for Jet Engine Designers

K-Search

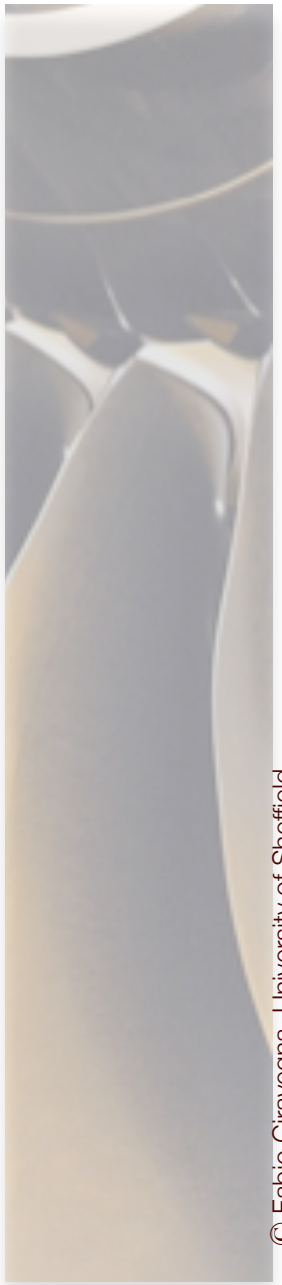


K-Search



K-Search evaluation

- We have performed 2 types of evaluations using K-Search:
  - in vitro:
    - Effectiveness of query strategy with respect to standard KS and OS
  - in vivo: testing the system with real users
    - 32 users Rolls-Royce engineers
      - Evaluation enables verifying suitability for use in a real environment



Annotating Documents

- Automatic extraction of information from event report
  - 18,000 documents analysed
    - Mainly Forms implemented in Word
- Metadata generated according to an ontology developed by Aberdeen U
- Automatic extraction of metadata and indexing of documents

IE unable to cover all the ontology with sufficient accuracy

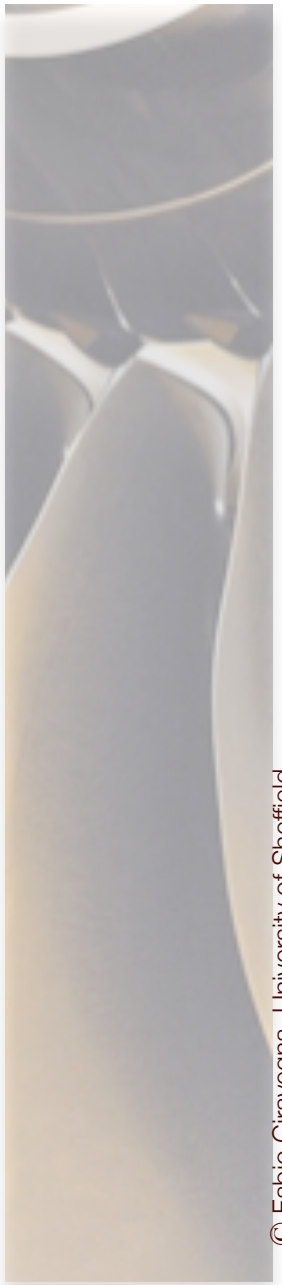
Applying information extraction

- AktiveMedia to annotate texts
- TRex system (Jiria et al. 2006) to train and extract
  - http://tyne.shef.ac.uk/t-rex/
- IE captures all the information in tables
  - 99% of the information captured (recall=99)
  - 98% of proposed information is correct (precision=98)

	POS	ACT	CORR	WRONG	MISSED	PREC	REC	F1
airport	120	120	120	0	0	100	100	100
has_airframe_cycles	104	104	104	0	0	100	100	100
has_airframe_hours	104	104	104	0	0	100	100	100
has_author	120	120	120	0	0	100	100	100
has_engine_serial_number	120	120	120	0	0	100	100	100
has_engine_type	120	120	120	0	0	100	100	100
has_event_date	120	120	120	0	0	100	100	100
has_event_report_no	356	358	356	2	0	99	100	100
has_part_description_installed	120	113	111	2	9	98	93	95
has_part_description_removed	120	133	120	13	0	90	100	95
has_part_number_installed	120	113	111	2	9	98	93	95
has_part_number_removed	120	133	119	14	1	89	99	94
TOTAL	**1644**	**1658**	**1625**	**33**	**19**	**98**	**99**	**98**

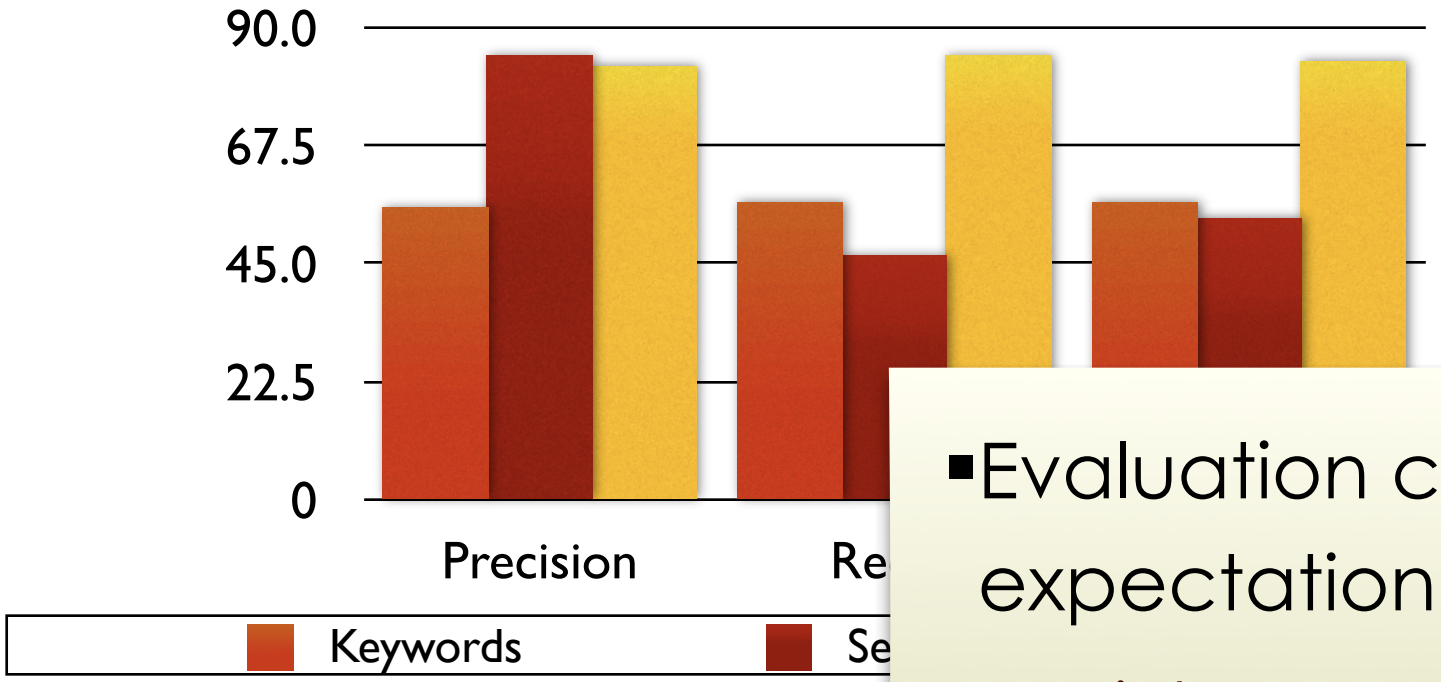
In vitro evaluation

- 21 topics of search, discussed with users, e.g.
  - "How many events were caused during maintenance in 2003?"
  - "What events were caused during maintenance in 2003 due to control units?"
  - 'Find al l the events associated with damage to acous- tic liners fol lowing bird strike"
- Queries:
  - "what events caused during maintenance in 2003 were due to control units?"
- Translated into a set of queries in KS, OS and HS



K-Search on Event Reports

- Accuracy in the first 20 hits on a sample of 400 docs



- Similar results for 50 hits

- Evaluation confirms our expectation:
  - Higher recall wrt OS and KS
  - Higher precision wrt KS
  - Slightly lower precision wrt OS

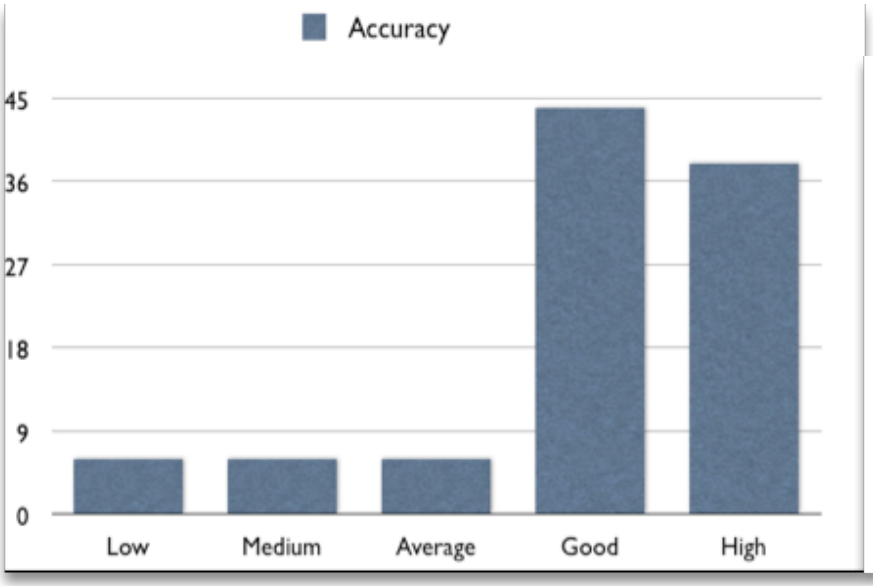
Final User Evaluation

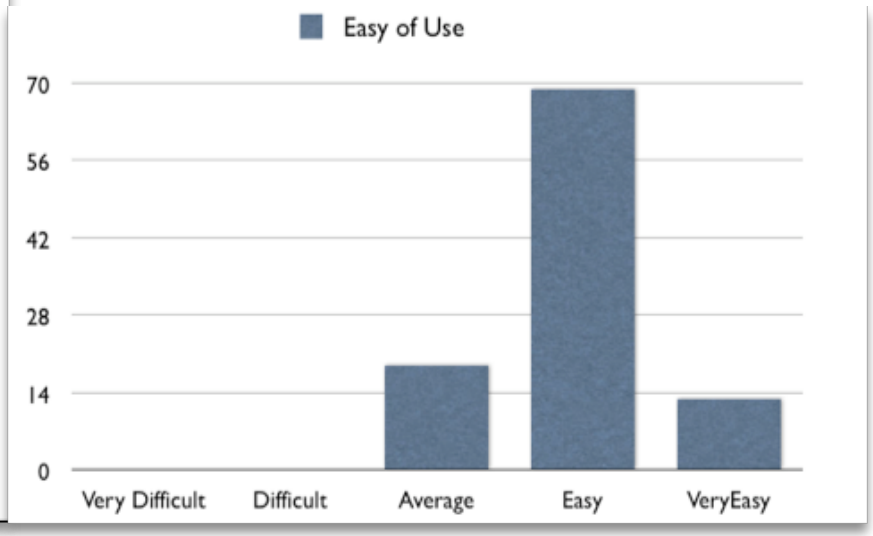
- Goal: verifying suitability for use in a real environment
  - 32 users Rolls-Royce engineers from different parts of the company
  - 90 minutes of test
    - Short introduction
    - 3 monitored tasks
      - One given (including solution)
      - One given (no solution)
      - One free task
  - Availability of system on intranet for the following period
- Evaluation: video recording, interview + log analysis

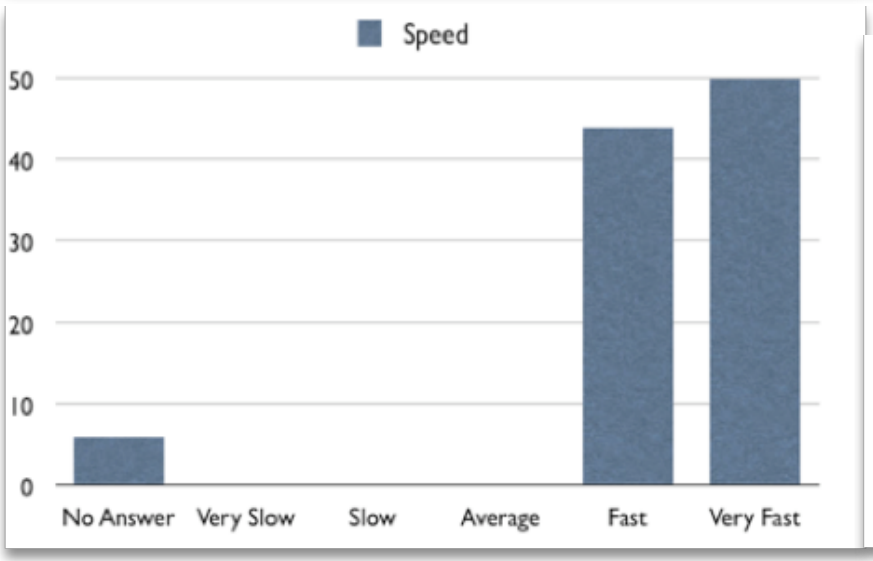
Evaluation Questions

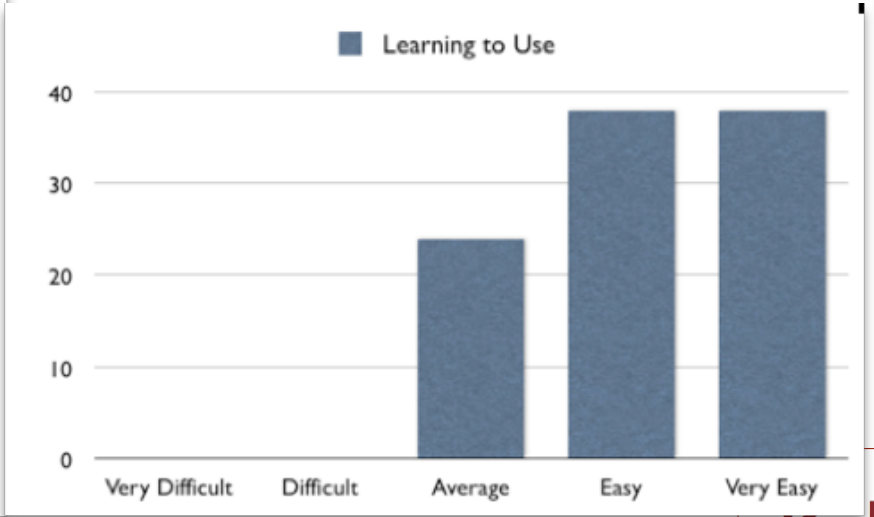
- Do user understand the hybrid paradigm?
- Are they able to search using HS?
- Do they actually use HS when confronted with a real searching task?
- Would the users be willing to use the system for their everyday work?

Liked by the users?

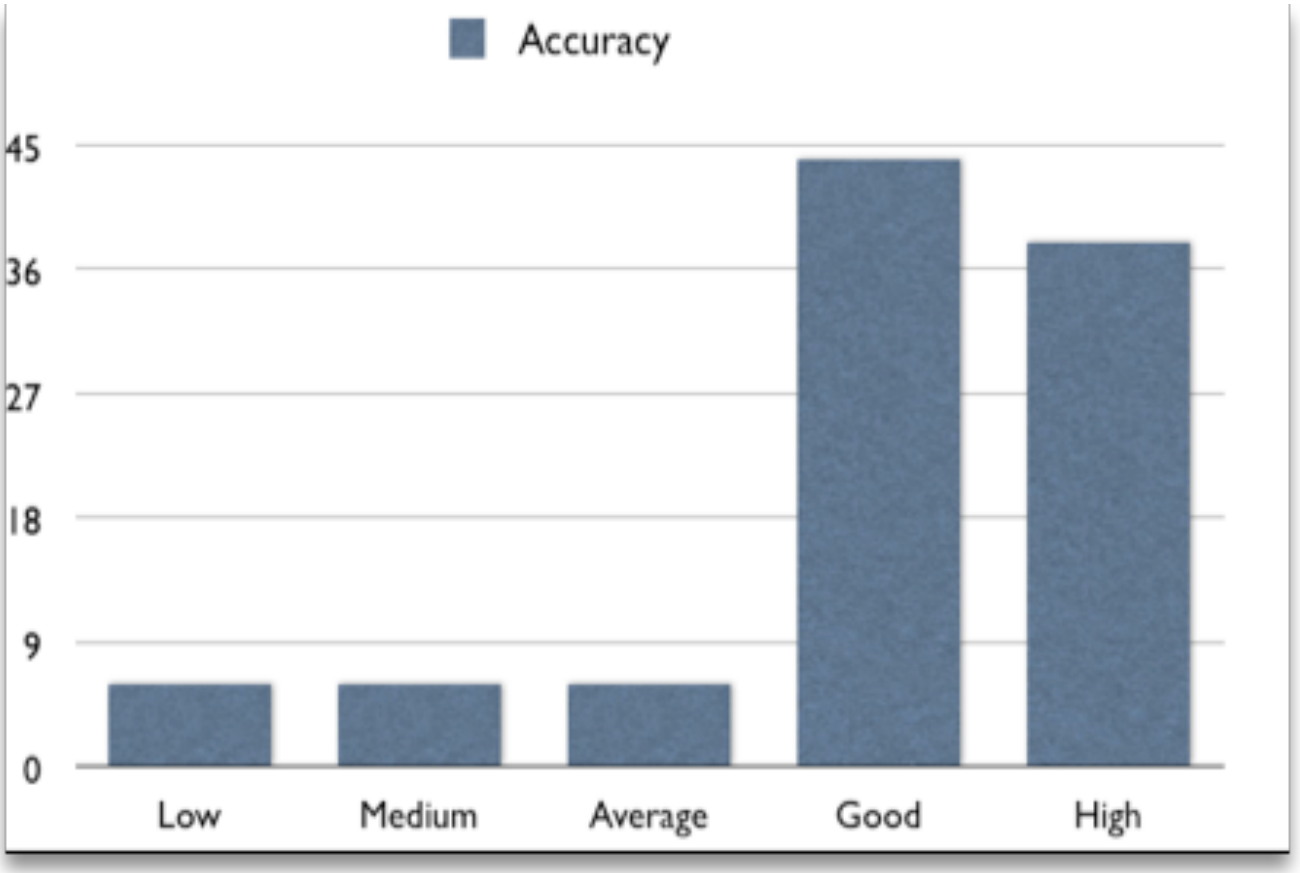




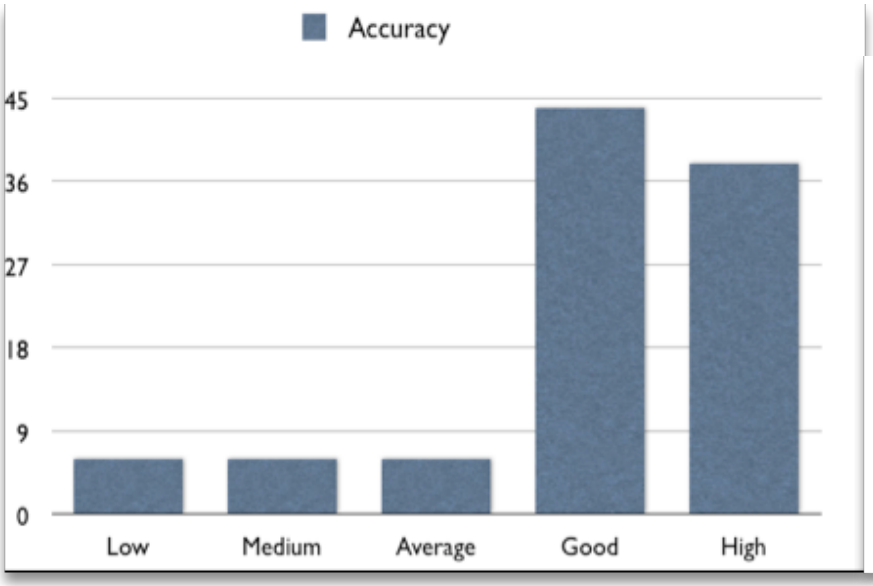


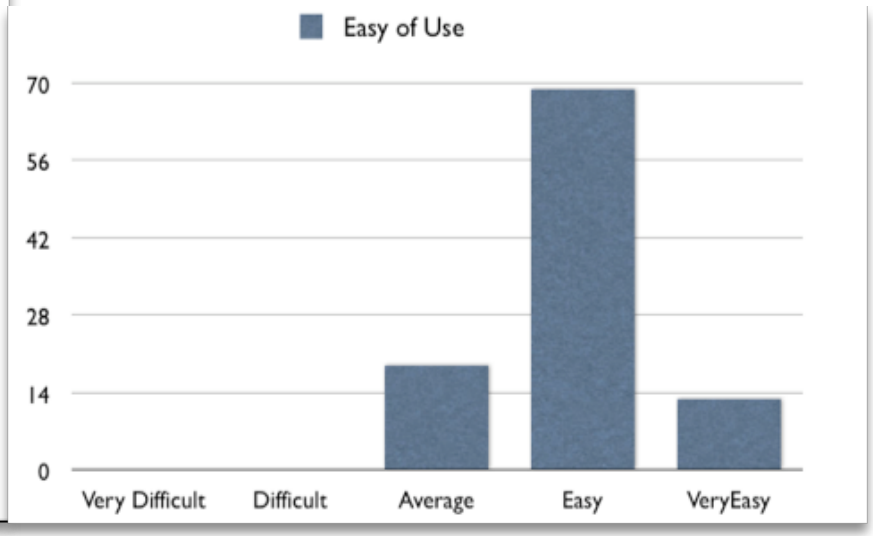


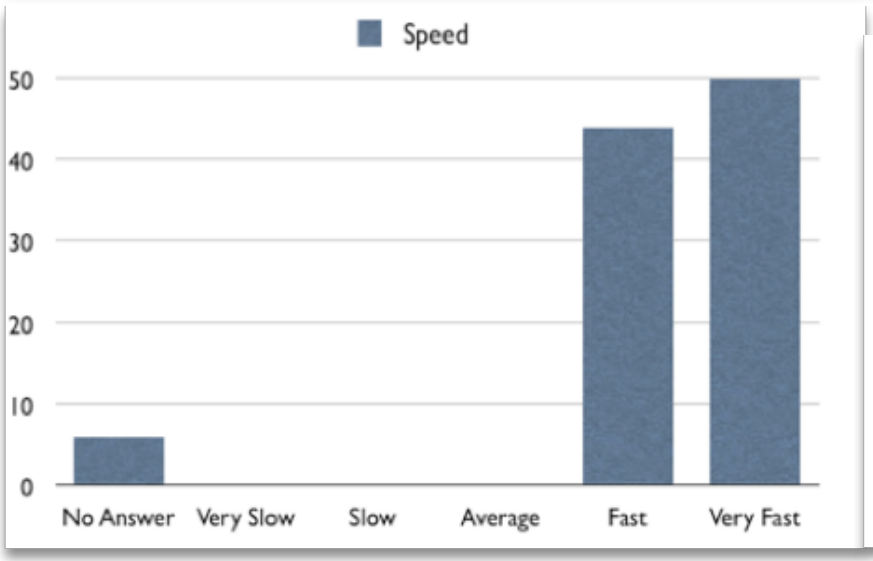
Liked by the users?

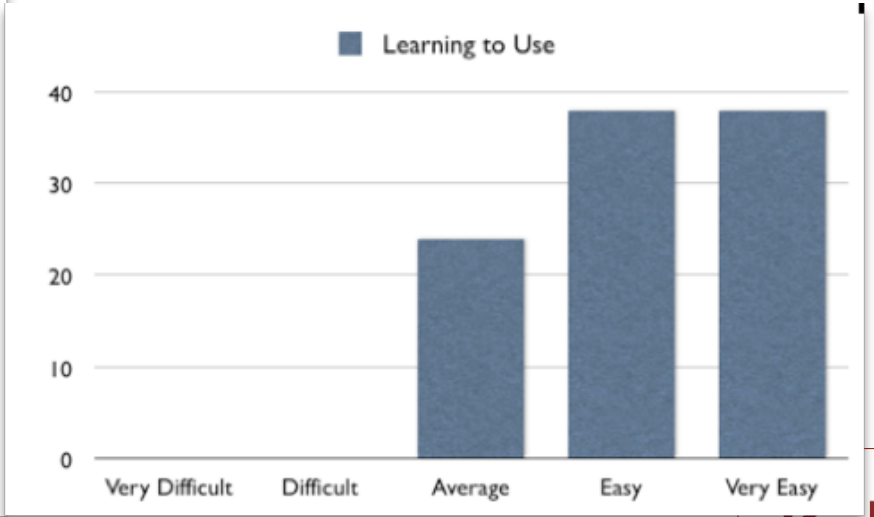


Liked by the users?

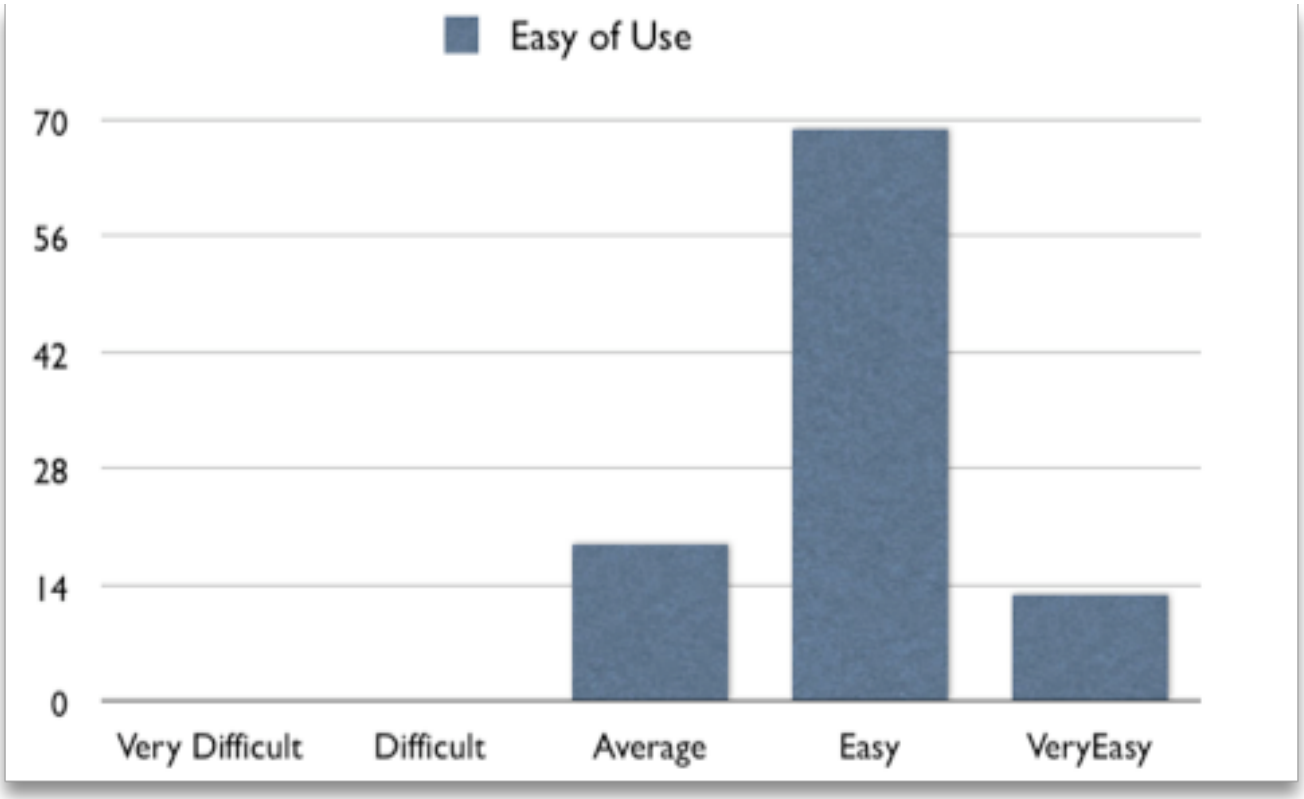




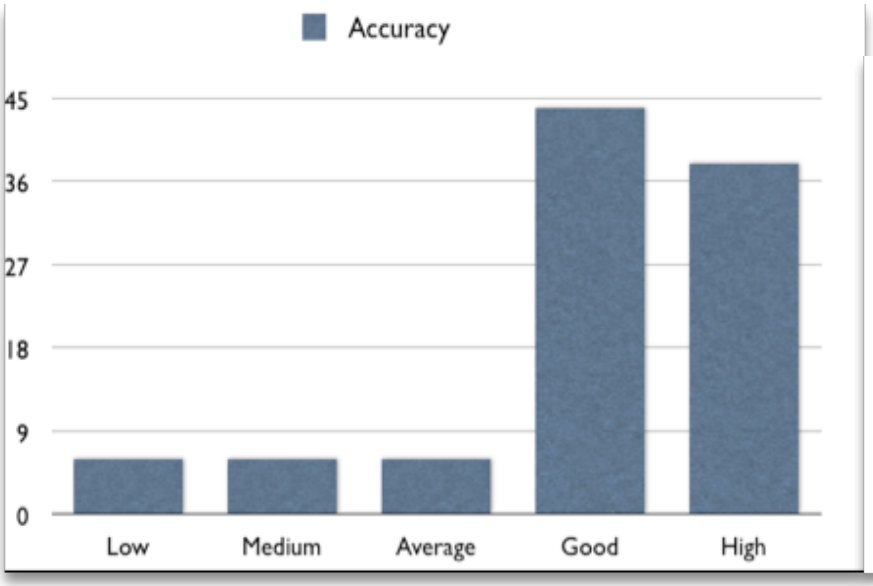


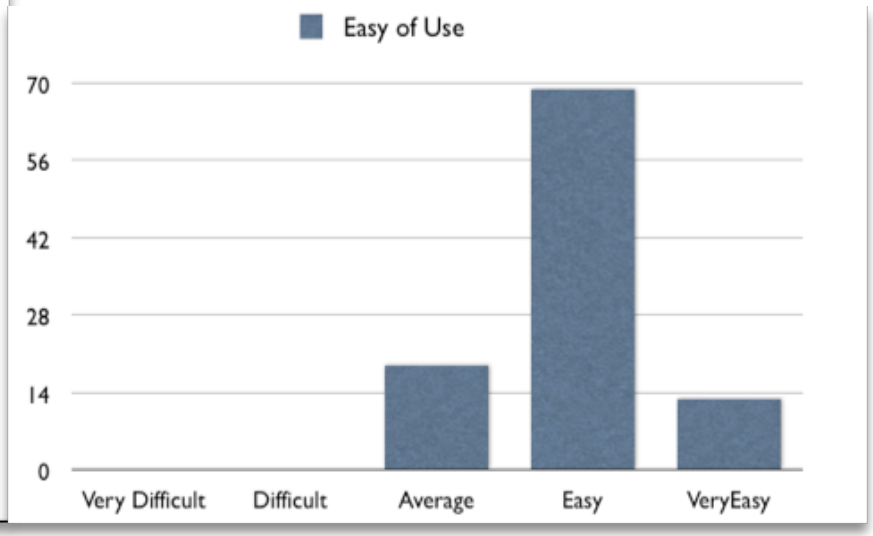


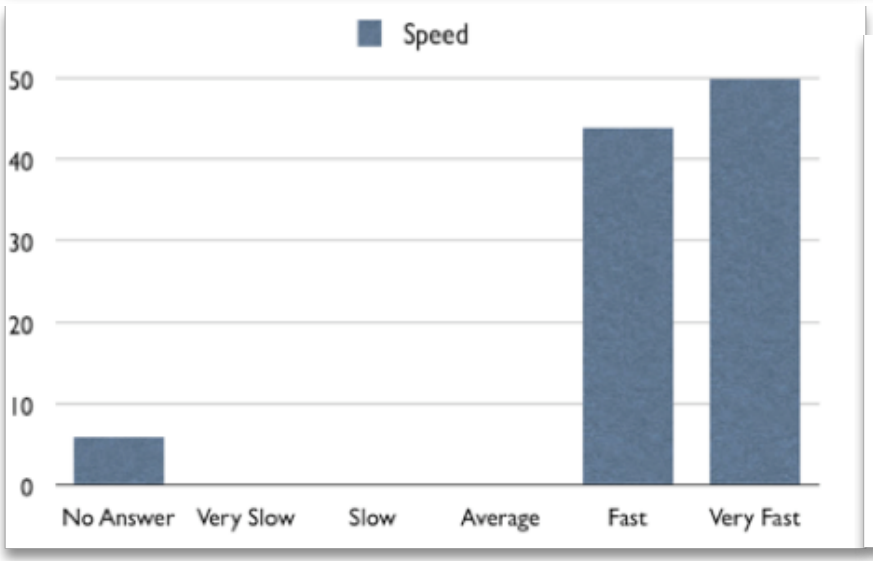
Liked by the users?

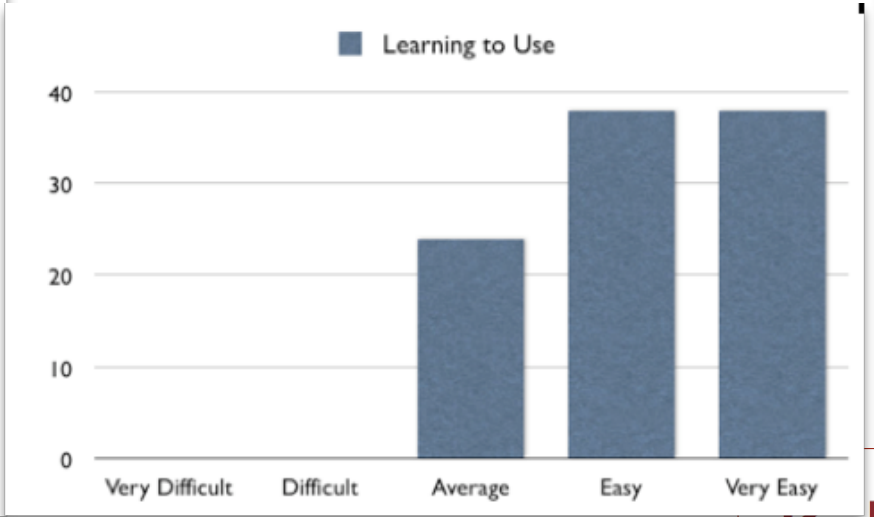


Liked by the users?

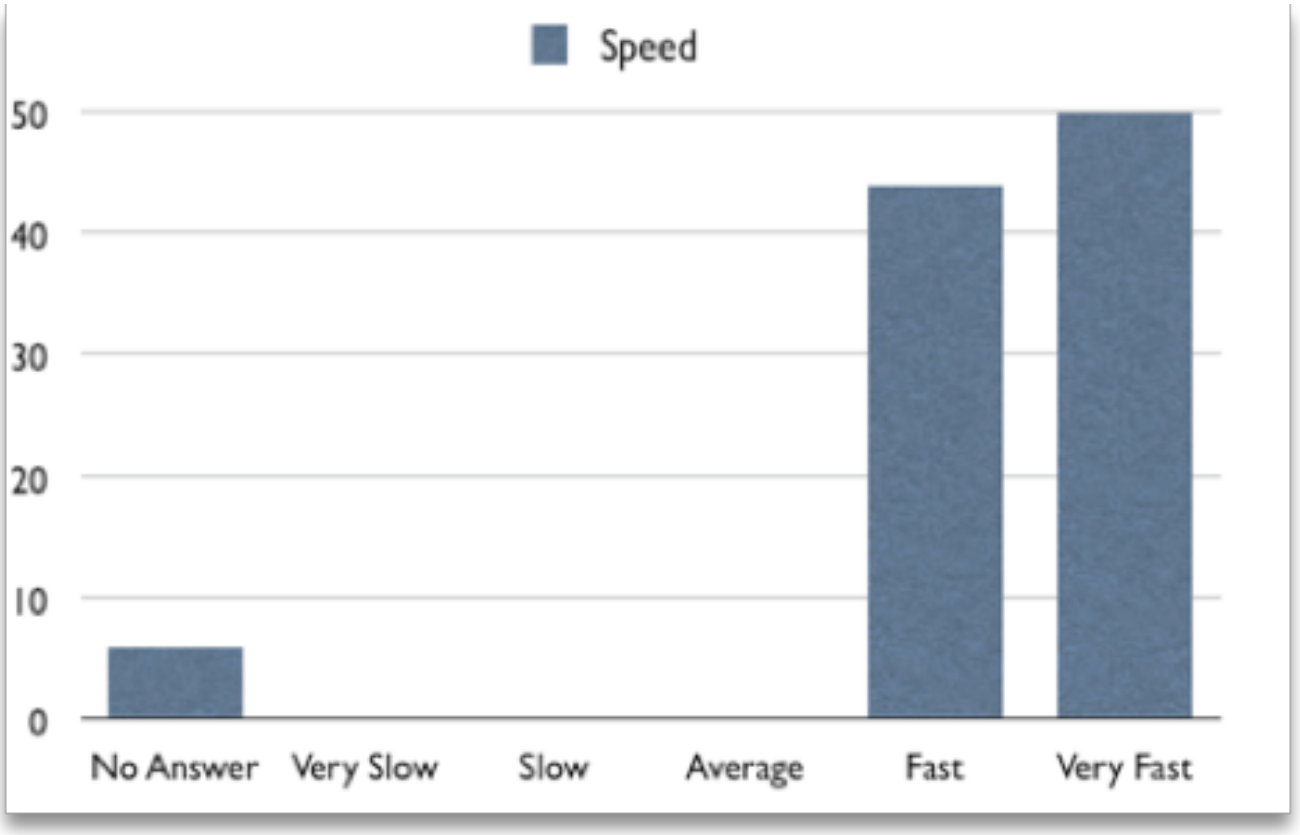




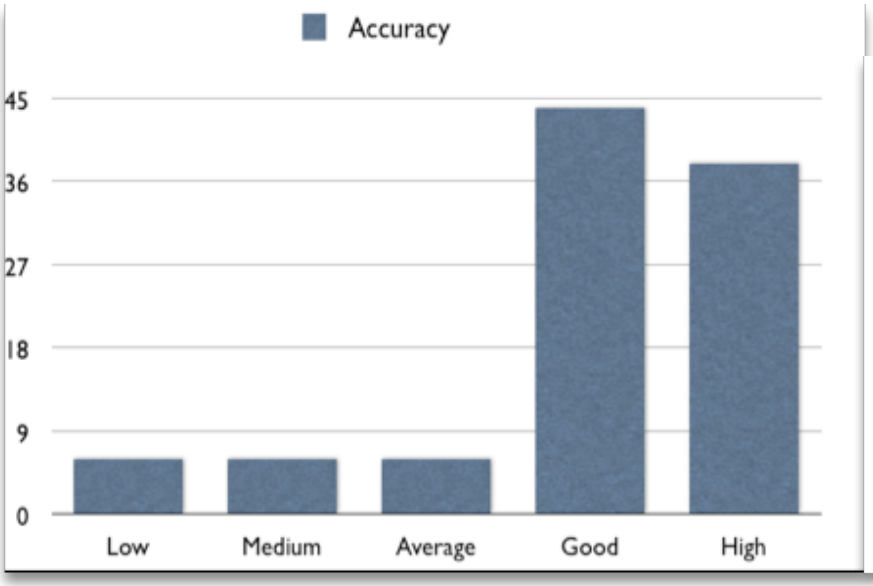


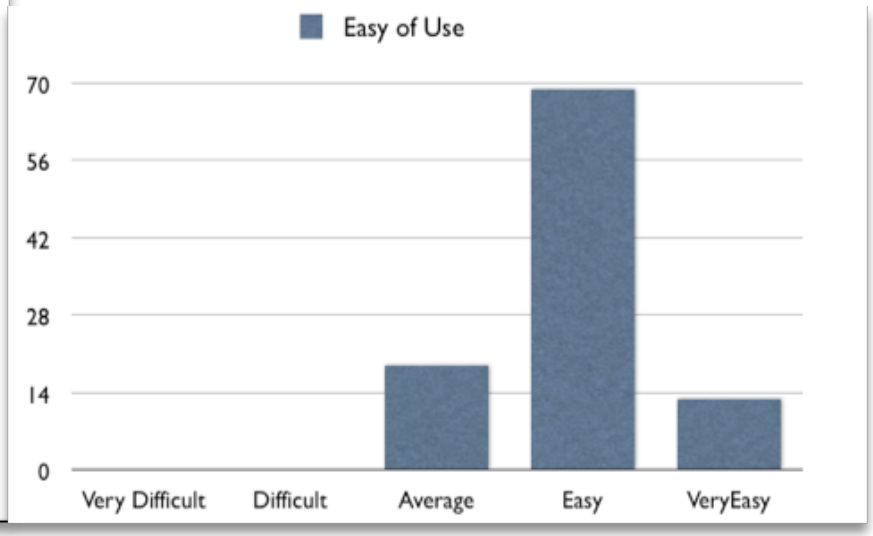


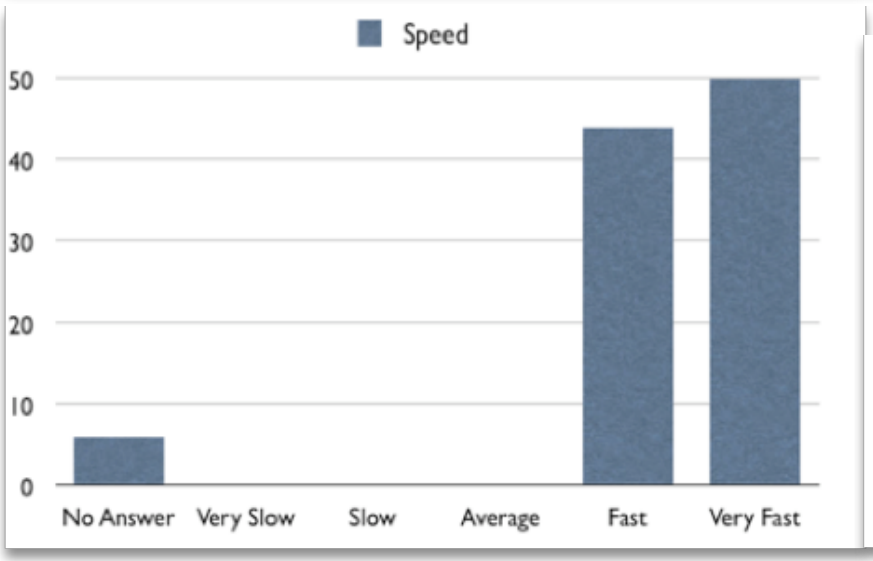
Liked by the users?

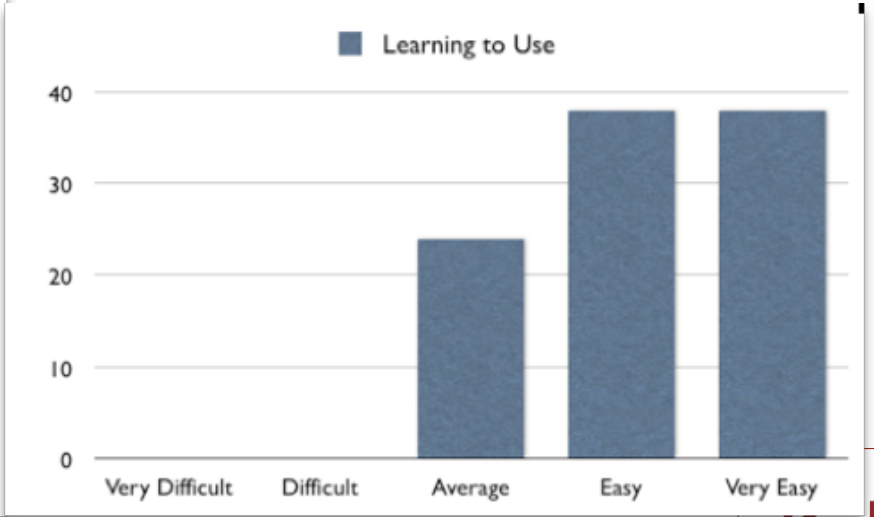


Liked by the users?

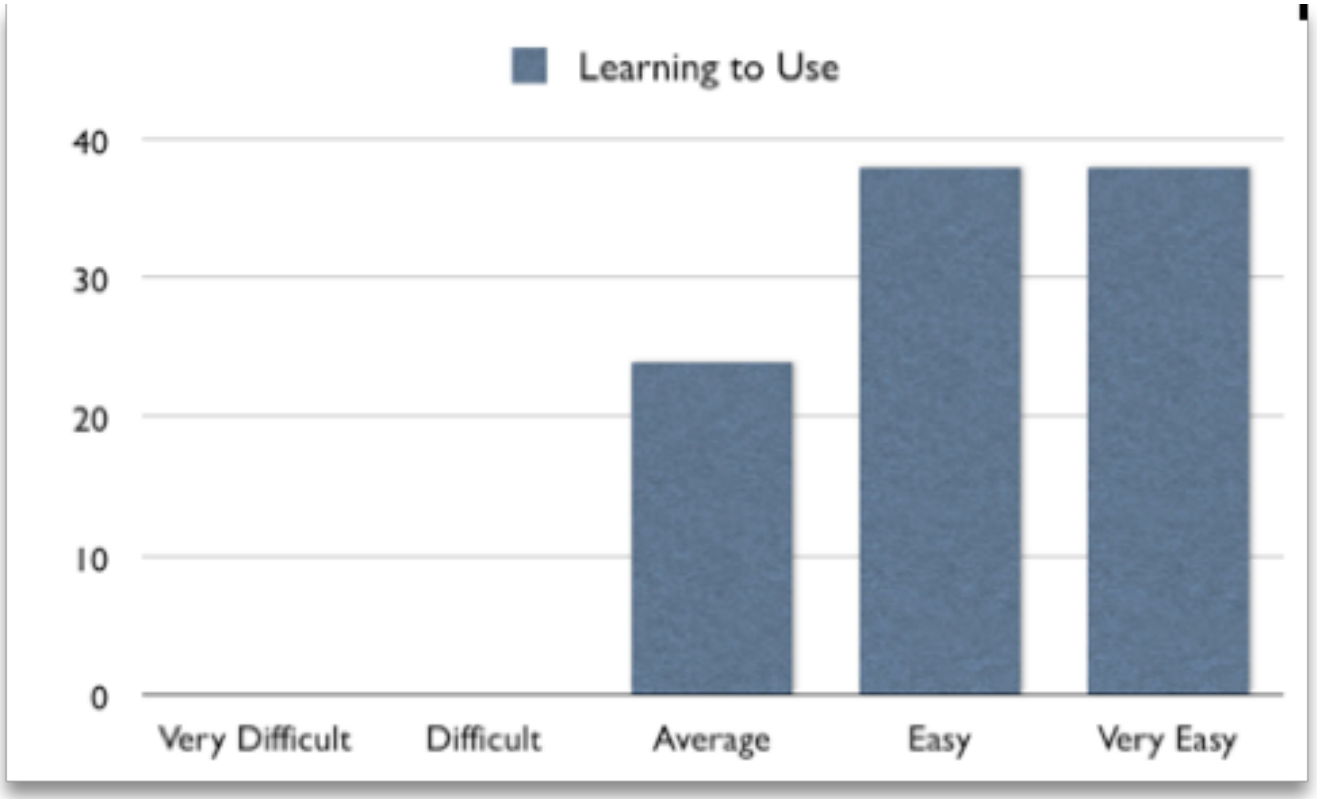




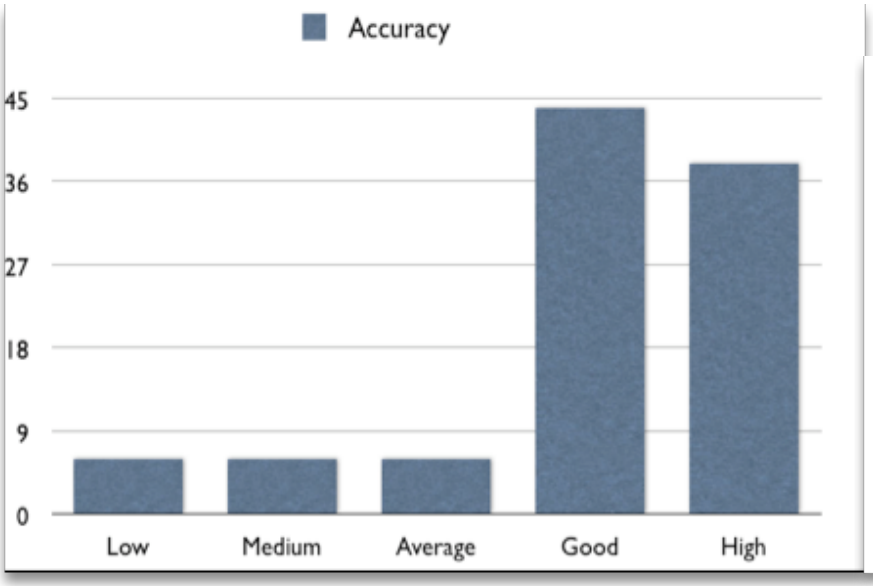


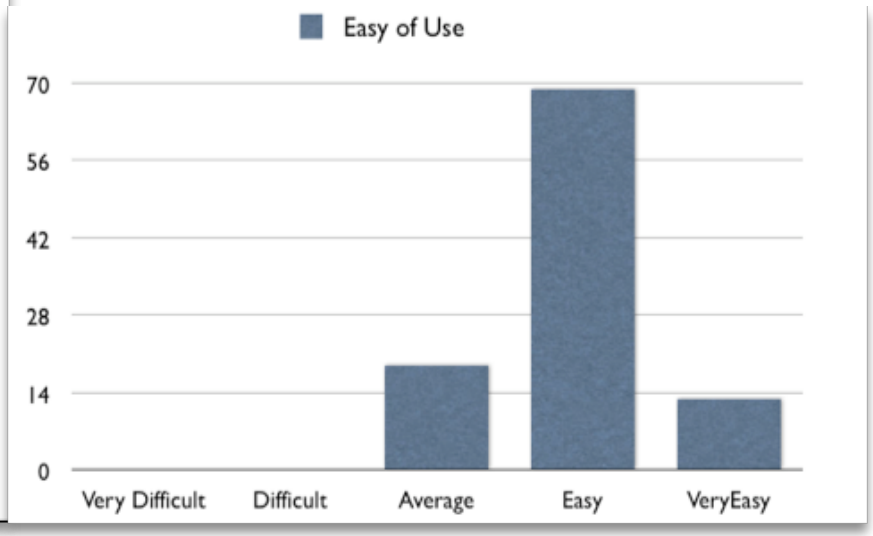


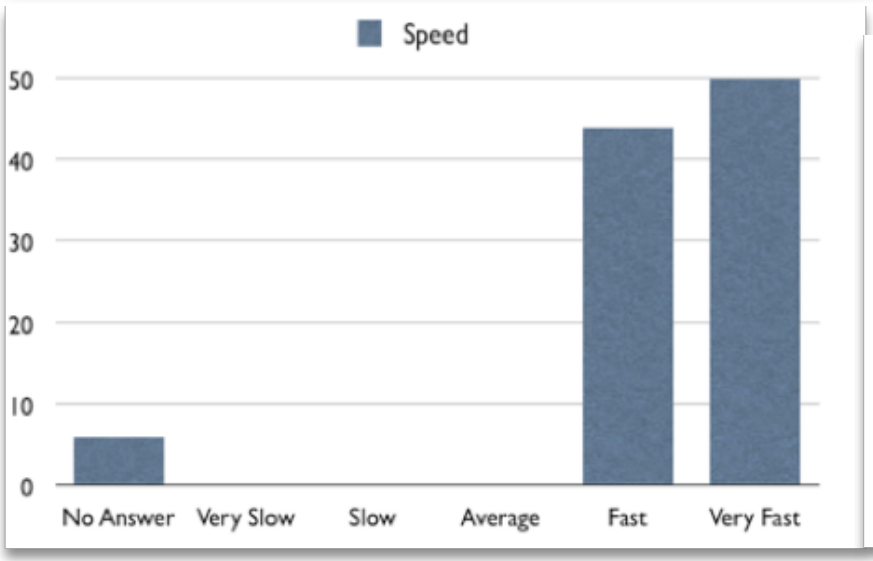
Liked by the users?

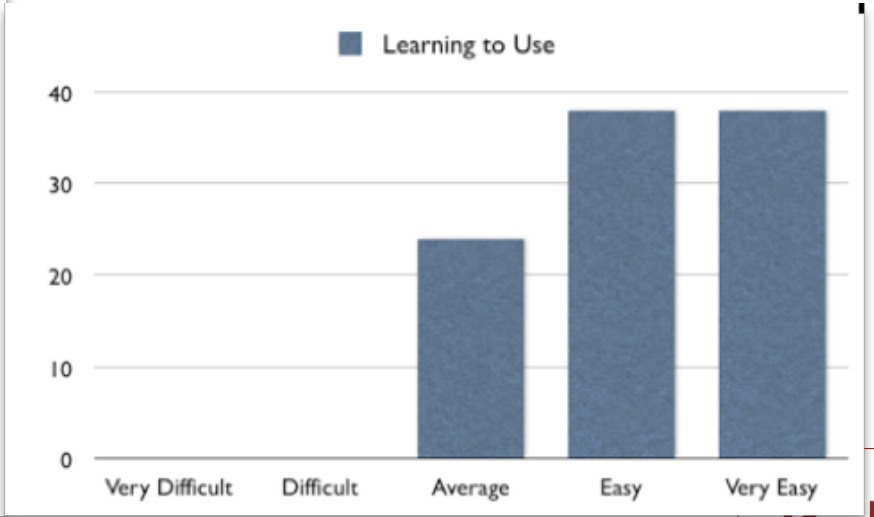


Liked by the users?









Search preferences: Service Engineers

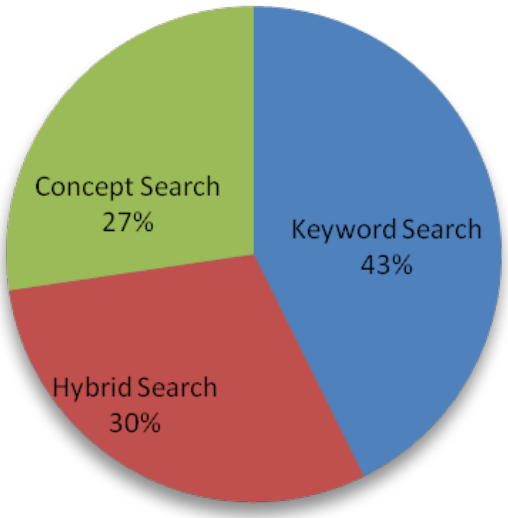
- Service engineers showed a clear predilection for hybrid search:
  - 61% of the search were executed using the hybrid modality
  - 24% using semantic search
  - 15% using keyword search.

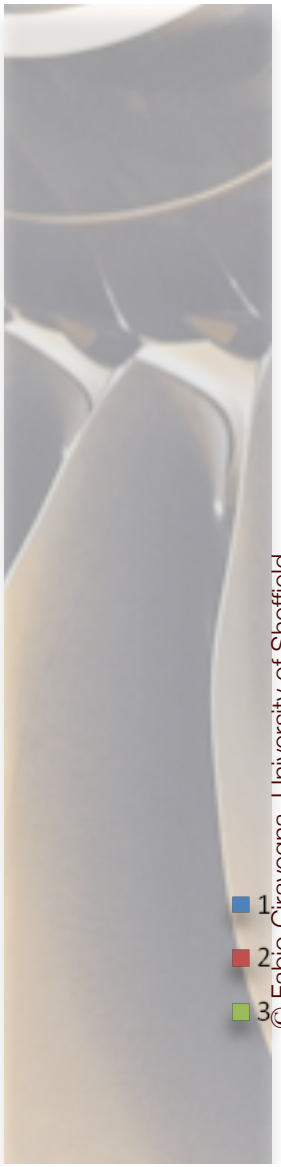
Go quickly on slides: just say different people used different strategies

Reason: data they were looking for was not all covered by the metadata

Search preferences: designers

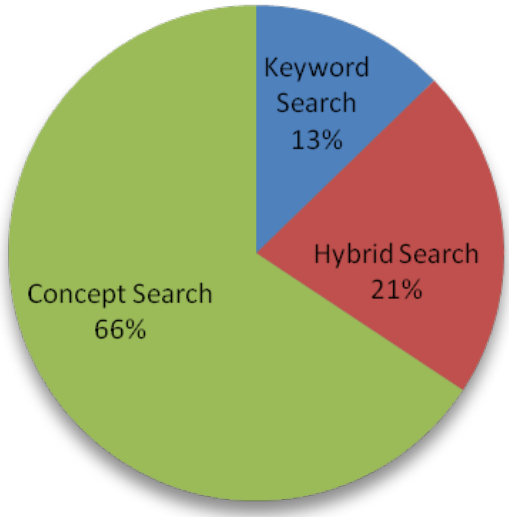
- Designers tended instead to favour keyword search:
  - 43% of the searches were executed using keyword search
  - 30% using hybrid
  - 27% using semantic search.





Search Strategies: Others

- The users belonging to other groups showed a predilection for concept search:
  - 66% of the searches were executed using semantic search
  - 24% using hybrid
  - 15% using keyword search.



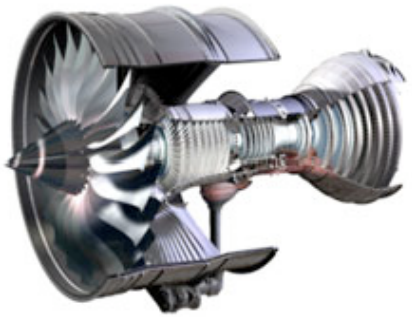
Liked by Users?

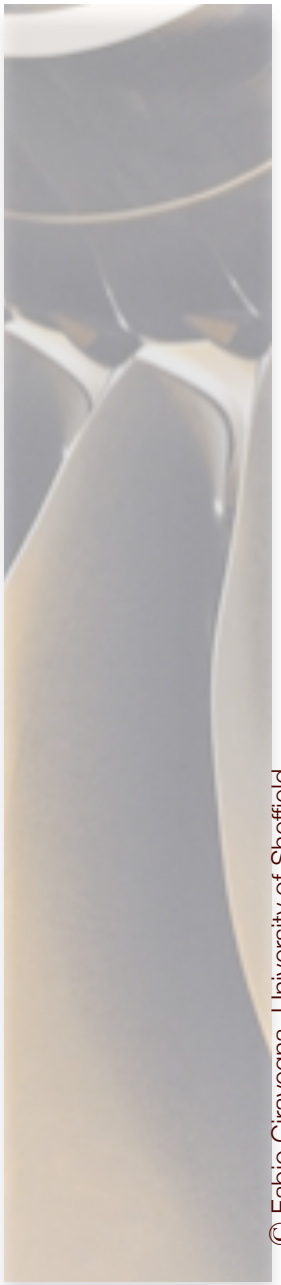
- Finalist of Rolls-Royce Director's Creativity Award 2007
  - Voted by employes for its innovation potential

Liked by Users?

- Support to the design of new Trent XWB
  - Porting to 9 Information Sources
    - 2008-2009
  - Carried out by:
    - 50% University
    - 50% k-now ltd (university spinout-company)

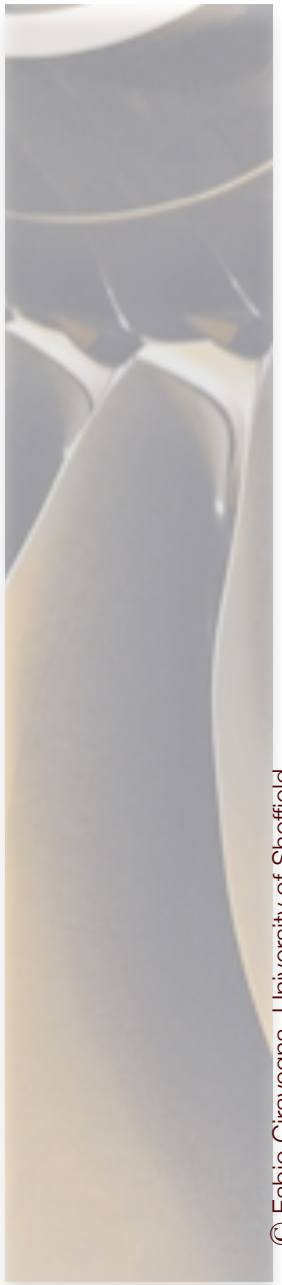
- Funds requested to UK Government for use of K-Tools for use in manufacturing





Conclusions

- Hybrid Search
  - It is compatible with the most used semantic search paradigms
    - Overcomes limitation of most current approaches based on metadata only
  - Accommodates different search strategies
    - Users can choose how to perform the query
  - Experimentally definitely outperforming both KS and OS



Future work

- Search across linked ontologies over intranet
- New ways of capturing information
  - User centred for new data
    - Cross-media
      - K-Forms
  - IE for legacy data
    - Cross-media

Thank You!

- Contact Information
  - www.dcs.shef.ac.uk/~fabio
  - fabio@dcs.shef.ac.uk
- Intelligent Web Technologies Lab
  - http://nlp.shef.ac.uk/wig/
- NLP Sheffield
  - http://nlp.shef.ac.uk/
- University of Sheffield
  - www.shef.ac.uk
- K-Now Ltd
  - www.k-now.co.uk