

Query log analysis and individual differences

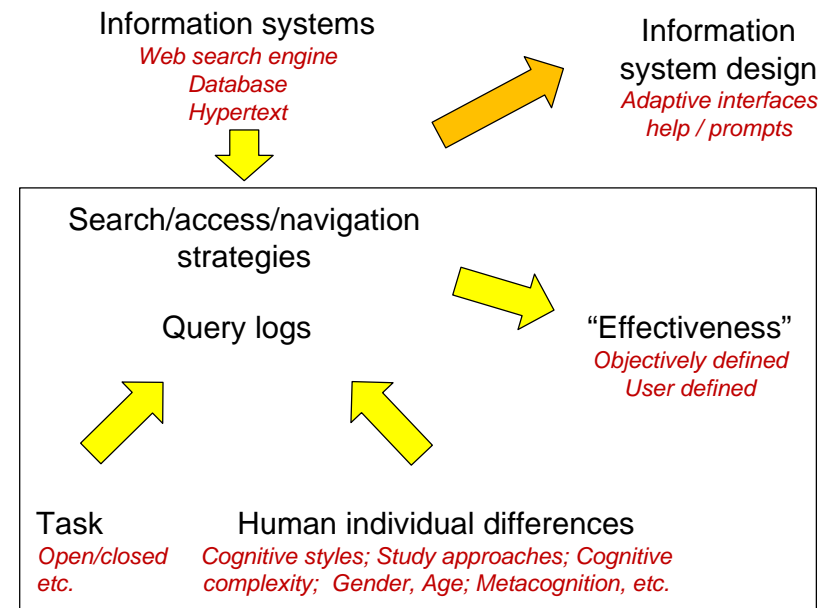
Nigel Ford
Department of Information Studies
University of Sheffield

Outline

- Type of projects carried out
- Techniques used
- Problems encountered
- Main limitations
- Likely future directions

Type of projects carried out

- We've been using logs to try to discover
 - strategic differences in the way different types of individual query information systems
 - the effects of any such differences on outcomes
 - The effects of system interventions on both the above (ongoing)



Type of projects carried out

- A couple of examples to give a flavour...

We started out...

- Wanting to observe how people behave in an “ideal” situation
 - How would people respond to an “ideal” information system?

An “ideal” system

- A database that could handle any question, no matter how phrased
- Was an expert in its subject area

An “ideal” system

- The database containing information on a document indexing system
- Students were tasked with learning about the document indexing system by interrogating the database in any way they liked
- They were tested on their resultant knowledge

An “ideal” system

- It was of course a Wizard of Oz experiment
 - Consisted of 2 on-line human experts...
 - backed up by computer files/ documentation
 - All interactions were logged

An “ideal” system

- Responded to questions, thus forcing students to be active
- Not too forthcoming, in order to encourage users to adopt some form of strategy

Level of questions

- The subject content was analysed into 8 hierarchical conceptual levels

Type of questions

- The query logs were analysed, and each question was inductively classified according to its intention
 - Descriptive
 - Focusing
 - Concrete
 - Analytic

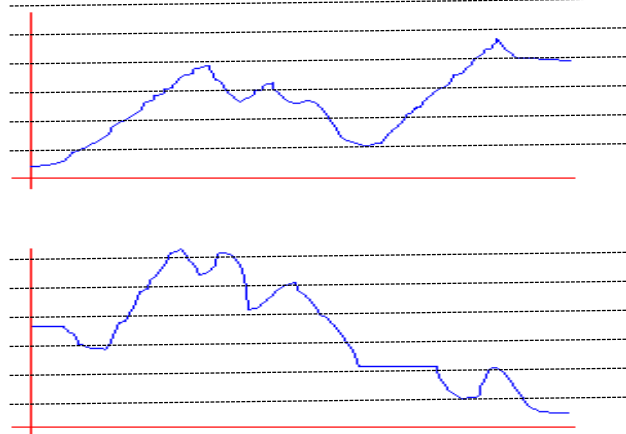
Type of questions

- Descriptive
 - Invited a straightforward descriptive answer
- Focusing
 - Sought to delimit / explore the bounds of a concept
- Concrete
 - Practical request to see some aspect of the indexing system in action
- Analytic
 - Reflecting a degree of analysis (e.g. to verify the searcher's understanding) of a concept)

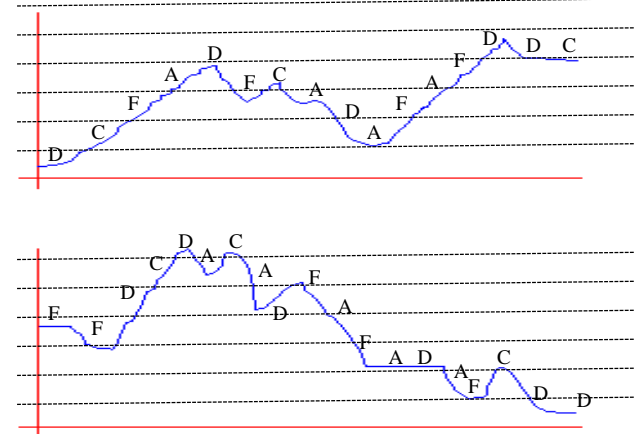
Data (for each student)

- Number of questions asked
 - ...at each hierarchical level
 - ...of each question type
- Test result (simple recall of information)
- Gender

Logs of learning strategies



Logs of learning strategies



Different strategies

- Were identified
 - “Deependers”
 - Dived in quickly to explore information deep in the subject hierarchy
 - “Mid-shallowenders”
 - Concentrated much more on higher levels in the subject hierarchy
 - “Consolidators”
 - A mix of the above – took longer to get to deep levels, but did explore them

Successful strategy 1

- Relatively passive intake of information (many Descriptive questions / few Analytic)
- Focusing on procedural detail at low levels in the subject hierarchy
- Female students

Successful strategy 2

- Relatively active approach (few Descriptive / many Focusing questions)
- Concentration on concepts relatively high in the subject hierarchy
- Male students

Cognitive styles

- These strategies mapped well onto an existing well established dimension of cognitive style (global/analytic).

Cognitive styles and web searching

- We focused in on cognitive styles and possible effects on web searching

Web searching

- 91 members of the general public performed 195 web searches
- Queries logged
- Cognitive styles measured
- More detailed analysis of the query logs

Searches

- Find the postcode of the tallest British building outside of London.
- You've received a postcard from friends who say they are abroad, visiting somewhere called Map. Where are they?
- There are many opportunities to win things on the Internet. Find some that may be of interest to you.
- What was written on Neville Chamberlain's piece of paper?
- You have won a trip to Saga. Can you find out anything interesting about the place?

Volunteer ID	QueryNo	Auto_SS	Query
011105mw01	7 *		tallest British building-London
011105mw01	8		tallest British building -London
011105mw01	9		tallest British building -London
011105mw01	10		tallest British building -London
011105mw01	11		2000 tallest British building -London
011105mw01	12		2000 tallest British building -London
011105mw01	13		2000 tallest British building -London
011105mw01	14		2000 tallest British building -London
011105mw01	15		2000 tallest British building -London -Americ
011105mw01	16		2000 tallest British building -London -Americ
011105mw01	17		2000 tallest British building -London -Americ
011105mw01	18		2000 tallest British building -London -Americ
011105mw01	1 *		vietnam rail timetable
011105mw01	2		vietnam train timetable
011105mw01	3 *		voluntary teaching abroad
011105mw01	4 *		Neville Chamberlain
011105mw01	5 *		Saga
011105mw01	6		Saga
011105mw02	1 *		Tancredi Pasero
011105mw02	2		Tancredi Pasero + biography
011105mw02	3		Tancredi Pasero + biography
011105mw02	4		"Tancredi Pasero" + biography
011105mw02	5		biography of Tancredi Pasero
011105mw02	6		biography of Tancredi Pasero

<i>Reduce a query</i>	A disjoint modification, which reduces a previous query in which it is contained as a sub-phrase.
<i>Reuse part of a query</i>	A modified query that has a sub-phrase in common with a previous query; this sub-phrase is shorter than either query in question. The sub-phrase forms only part of the prior query.
<i>Reconfigure part of a query</i>	As above but the number of words in common is also greater than the word-length of the common sub-phrase: an indication of some re-ordering of words, word insertion or removal.
<i>Retain a single word</i>	A modified query with a single word in common. The single word forms only part of the prior query
<i>Retain separated words</i>	A modified query that has more than one word in common with a previous query but these are separated in one or both queries. Usually indicates an insertion or word replacement between common words.

<i>Reduce a query</i>	houses for sale in lews county	houses for sale
	space shuttle discovery	shuttle discovery
<i>Reuse part of a query</i>	photos of	photos of rio de janerio
	biography of elie weisel	picture of elie weisel
<i>Reconfigure part of a query</i>	medical etiquette – use of credentials	use of nursing credentials
	Putnam County Illinois election results	election 2001 results for Putnam County Illinois
<i>Retain a single word</i>	plumbing sales	toilet sales
	“Pneumococcol Bacteriamia”	“Pneumococcus Bacteriamia”
	rowing footwear	jl design rowing
<i>Retain separated words</i>	USD skate pictures	USD 2000 team skate
	motor rewinding	motor specifications rewinding
	“cours de gestion financière”	“cours d’analyse financière”

Query Transformations Primary codes

Code	Query Transformation
U	A unique query. Only used for a single query session.
N	A new query (recognised by being at the start of a session or having low textual similarity to preceding queries) appearing during a session of at least two queries.
R	A repeated query with the same page rank - probably seeking relevance feedback.
P	A repeated query with increased page rank - further investigation of results from the current query.
p	A repeated query with reduced page rank - further investigation of results from the current query by returning to earlier pages.
I(k)	Indicates an identical query (including quotation marks and Boolean operators) to the one in the k'th position. This excludes identical queries in the immediately preceding position, which are covered by codes 'R', 'P' and 'p' in Table 1.
J(k)	Indicates an identical query apart from quotation marks, and/or Boolean + marks.

Code	Query Transformation
C(k)	A conjoint modification, which extends query k and retains it as a sub-phrase.
D(k)	A disjoint modification, which reduces a query k in which it is contained as a sub-phrase.
S(k)	A modified query that has a sub-phrase in common with query k; this sub-phrase is shorter than either query in question. The sub-phrase forms only part of the prior query.
s(k)	As S(k) above but in this case the number of words in common is also greater than the word-length of the common sub-phrase: an indication of some re-ordering of words, word insertion or removal.
W(k)	A modified query with a single word in common with query k. The single word forms only part of the prior query
w(k)	A modified query that has more than one word in common with query k but these are separated in one or both queries. Usually indicates an insertion or word replacement between common words.
M(k)	A modified query recognised on the basis of some textual similarity with a previous query above the threshold level. It cannot be further categorised as one of the above and probably contains changed word

Code	Query Transformation
Z(k)	Queries not recognised as similar but found to have a single word in common with query <i>k</i> . This word is probably short, in comparison with the query length; frequently 'and' 'of' or 'in'.
z(k)	Queries not recognised as similar but found to have more than one word (as above, usually short) in common with query <i>k</i> .

Query Transformations Supplementary Codes

Code	Query Transformation
B	Indicates the inclusion of a Boolean operator (+, AND, OR, NOT).
b	Indicates the removal of a Boolean operator.
Q	Indicates the inclusion of quote marks.
q	Indicates the removal of quotation marks
_	Delay term indicating prolonged inactivity prior to a subsequent query.

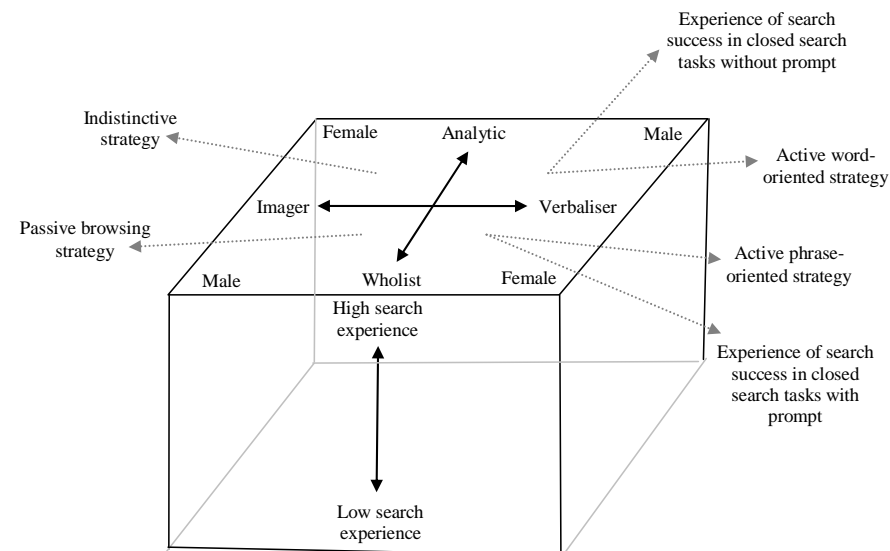
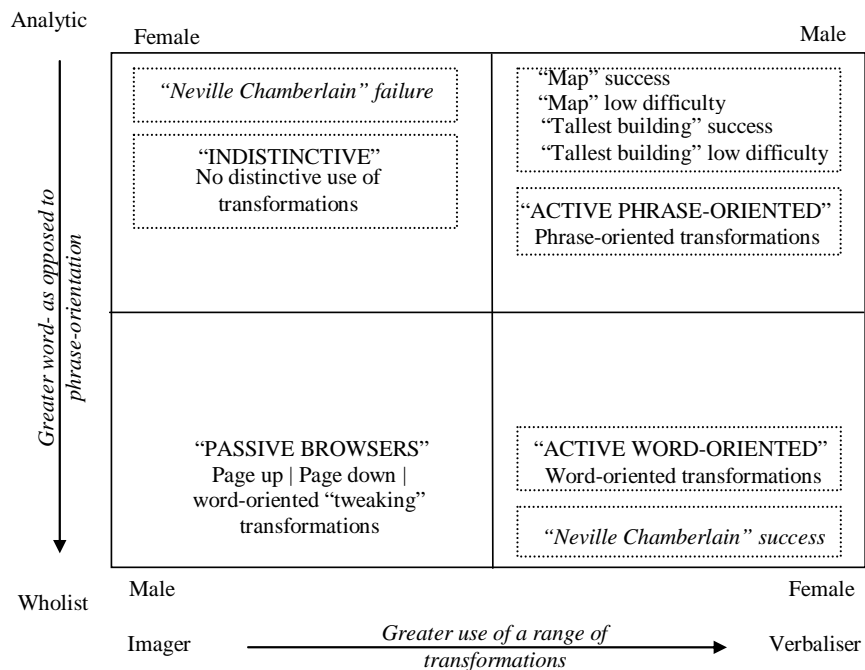
Volunteer ID	QueryNo	Auto_SS	Query	QM(similarity)
011105mw01	7 *		tallest British building-London	BN
011105mw01	8		tallest British building -London	J(7)
011105mw01	9		tallest British building -London	R
011105mw01	10		tallest British building -London	P
011105mw01	11		2000 tallest British building -London	C(7)
011105mw01	12		2000 tallest British building -London	P
011105mw01	13		2000 tallest British building -London	R
011105mw01	14		2000 tallest British building -London	P
011105mw01	15		2000 tallest British building -London -Americ	BC(11)
011105mw01	16		2000 tallest British building -London -Americ	R
011105mw01	17		2000 tallest British building -London -Americ	P
011105mw01	18		2000 tallest British building -London -Americ	R
011105mw01	1 *		vietnam rail timetable	N
011105mw01	2		vietnam train timetable	w(1)
011105mw01	3 *		voluntary teaching abroad	N
011105mw01	4 *		Neville Chamberlain	N
011105mw01	5 *		Saga	N
011105mw01	6		Saga	R
011105mw02	1 *		Tancredi Pasero	N
011105mw02	2		Tancredi Pasero + biography	BC(1)
011105mw02	3		Tancredi Pasero + biography	R
011105mw02	4		"Tancredi Pasero" + biography	QJ(2)
011105mw02	5		biography of Tancredi Pasero	bs(2)
011105mw02	6		biography of Tancredi Pasero	R

Typical results

- *Verbalisers* display the most extensive distinctive use of search transformation strategies..
- *Analytic* verbalisers are characterised by *phrase-oriented* searching, (entailing “Add to a query”, “Reuse part of a query”, “Reconfigure part of a query”. These are all phrase-oriented rather than word-oriented, as shown by their definitions which all require the presence of phrases.

Typical results

- *Wholist* verbalisers, however, are characterised by *word-oriented* searching, entailing “Retain a single word”, “Retain separated words”, “1 word in common – otherwise dissimilar” and “Other modification”. These are all word-oriented, as shown by their definitions.



Problems & limitations

- **Complexity**
 - Complex results often not as predicted from theory
 - Not knowing which variables to include
- **Measurement**
 - Measures not well developed (e.g. cognitive styles)
- **Scale**
 - Need to gather personal data makes large scale investigations difficult
- **Coordination**
 - Complexity and measurement issues make more difficult large scale coordinated research

Future directions?

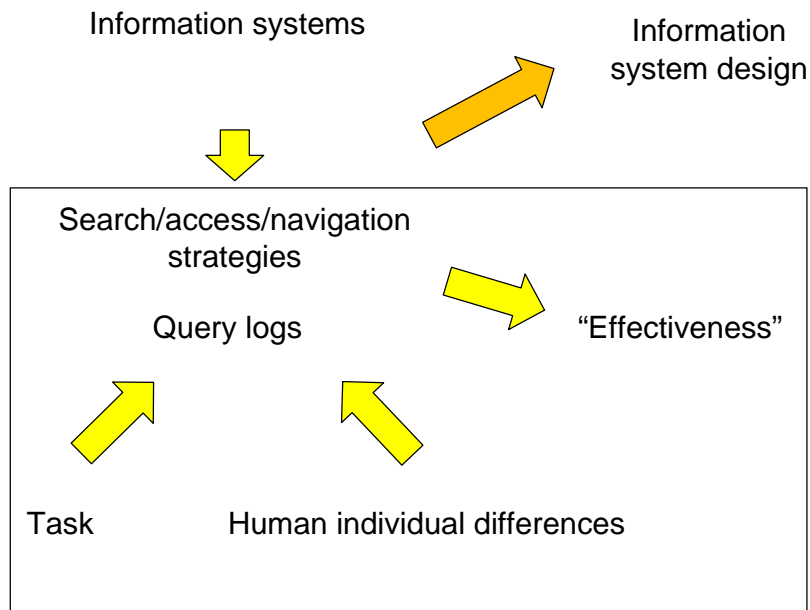
- More accurate and reliable – and agreed – measures of relevant variables (including human individual differences, search task characteristics, search behaviour, and search outcomes);
- Better identification of – and agreement on – appropriate variables that may be influential in the phenomena we wish to understand;
- Much larger data sets in order to enable modelling;
- More sophisticated modelling of complex non-linear relationships between variables.

But...

- It may be that, at highly complex levels of information behaviour, interactions between variables
 - relating to different people at different times and in different conditions, with different information needs, search behaviours, search outcomes and reactions to them over time
- are themselves too dynamic, variable, and susceptible to serendipitous interventions, to be subject to accurate and reliable measurement and modelling.



To conclude...

- The most significant illumination in our understanding of user behaviour may in the future derive from more powerful and sophisticated quantitative models based on large coordinated research efforts.
- However, it is also possible that such approaches may take us only to a certain plateau ...
- We need to use query log analysis to help develop smarter users as well as smarter systems.





Extras....

Cognitive styles test

Is this  the same as 



No Yes

Cognitive styles test

Is this  the same as 

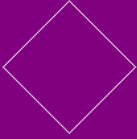

No Yes

Cognitive styles test

Is this  the same as 

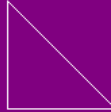

No Yes

Cognitive styles test

Is this  contained in 

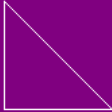
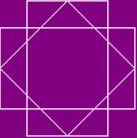
No Yes

Cognitive styles test

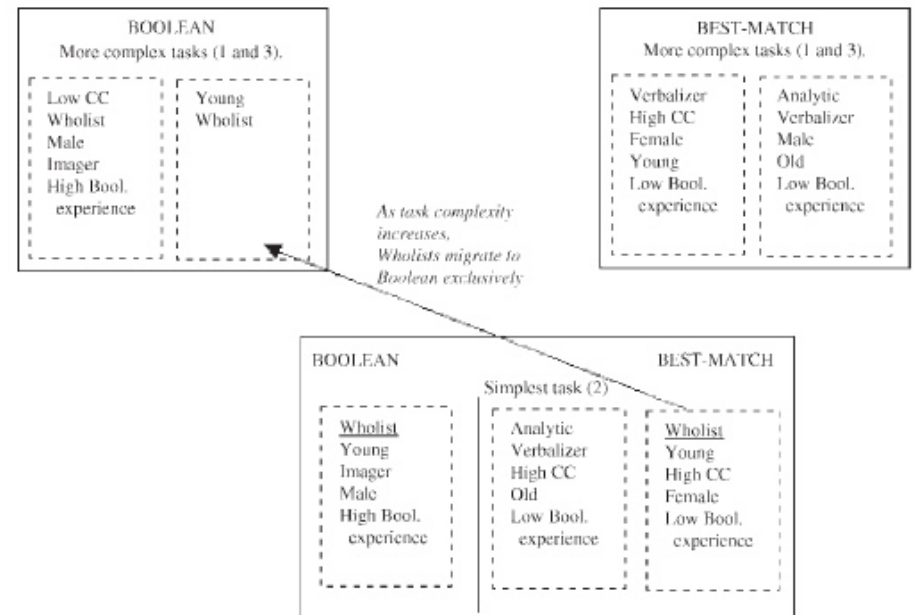
Is this  contained in 

No Yes

Cognitive styles test

Is this  contained in 

No Yes



Broad strategies

Global cognitive style

Aware of broadening/
narrowing techniques
Dissatisfied with search
results
Many different search
terms used
Many new search
terms used
Many relevant references
retrieved

Analytic cognitive style

Less aware of broadening/
narrowing techniques
Satisfied with search
results
Many different search
terms used
Few new search
terms used
Fewer relevant references
retrieved

Very simple log analysis

	Most complex		Least complex		Most complex		
	Task 3	Task 1	Task 2	Task 2	Task 1	Task 3	
Boolean							Best-match
Wholist	✓ (✓)	✓ ✓	✓	×	✓ ✓	✓	Analytic
Imager	✓	✓ ✓	✓	✓	✓ ✓	(✓)✓	Verbalizer
Low cog. complexity	✓	✓ ✓	✓	✓	✓ ✓	✓	High cog. complexity
Male	×	✓ ✓	✓	✓	×	×	Female
Young	✓	✓ ✓	✓	×	✓ ×	✓ ×	Old
High Boolean experience	✓	×	✓	(✓)✓	✓ ✓	✓ ✓	Low Boolean experience

Very simple log analysis

	Most complex		Least complex		Most complex		
	Task 3	Task 1	Task 2	Task 2	Task 1	Task 3	
Boolean							Best-match
Wholist	✓ (✓)	✓ ✓	✓	×	✓ ✓	✓	Analytic
Imager	✓	✓ ✓	✓	✓	✓ ✓	(✓)✓	Verbalizer
Low cog. complexity	✓	✓ ✓	✓	✓	✓ ✓	✓	High cog. complexity
Male	×	✓ ✓	✓	✓	×	×	Female
Young	✓	✓ ✓	✓	×	✓ ×	✓ ×	Old
High Boolean experience	✓	×	✓	(✓)✓	✓ ✓	✓ ✓	Low Boolean experience

uid 74:
NM(1)C(2)C(3)S(4)s(5)PPPPRRRRRRRpp(5)s(6)s(22)s(22)s(23)s(25)s(26)s(22)R

