adopting systematic evaluation benchmarks in operational settings

jussi karlgren
gavagai
spotify
kth royal institute of technology

these following points have been collected from

- discussions at many industrial events
- several CLEF labs
- projects CHORUS and PROMISE
- a workshop on “Practical Issues in Information Access System Evaluation”
- collected into a chapter in the CLEF book

these following points have been collected from

- discussions at many industrial events
- several CLEF labs
- projects CHORUS and PROMISE
- a workshop on “Practical Issues in Information Access System Evaluation”
- collected into a chapter in the CLEF book

these following points have been collected from

- discussions at many industrial events
- several CLEF labs
- projects CHORUS and PROMISE
- a workshop on “Practical Issues in Information Access System Evaluation”
- collected into a chapter in the CLEF book

these following points have been collected from

- discussions at many industrial events
- several CLEF labs
- projects CHORUS and PROMISE
- a workshop on “Practical Issues in Information Access System Evaluation”
- collected into a chapter in the CLEF book

these following points have been collected from

- discussions at many industrial events
- several CLEF labs
- projects CHORUS and PROMISE
- a workshop on “Practical Issues in Information Access System Evaluation”
- collected into a chapter in the CLEF book

these following points have been collected from

- discussions at many industrial events
- several CLEF labs
- projects CHORUS and PROMISE
- a workshop on “Practical Issues in Information Access System Evaluation”
- collected into a chapter in the CLEF book

“operational” settings = industrial, applied, both commercial and non-commercial etc

evaluation of information systems in operational settings is different

1. an information access service is usually not the primary objective
2. the objective is to perform some task adequately, not perfectly

1. an information access service is usually not the primary objective
2. the objective is to perform some task adequately, not perfectly

1. an information access service is usually not the primary objective
2. the objective is to perform some task adequately, not perfectly (optimisation beyond that is wasted development effort)

1. best practice

- customers have great readiness to accommodate to mediocre systems if aligned with their goals
- cost of introducing better systems may be prohibitive
- search is only one component in a system

1. best practice

- customers have great readiness to accommodate to mediocre systems if aligned with their goals
- cost of introducing better systems may be prohibitive
- search is only one component in a system

1. best practice

- customers have great readiness to accommodate to mediocre systems if aligned with their goals
- cost of introducing better systems may be prohibitive
- search is only one component in a system

1. best practice

- customers have great readiness to accommodate to mediocre systems if aligned with their goals
- cost of introducing better systems may be prohibitive
- search is only one component in a system

2. availability

many academic test sets are only available to non-profit or research organisations!

practical challenges

1. systems may have many instances, sometimes non-identical; usage may be distributed across numerous nodes;
2. the data under consideration may vary;
3. operational data can be messy, incomplete, and distributed over numerous systems

practical challenges

1. systems may have many instances, sometimes non-identical; usage may be distributed across numerous nodes;
2. the data under consideration may vary;
3. operational data can be messy, incomplete, and distributed over numerous systems

practical challenges

1. systems may have many instances, sometimes non-identical; usage may be distributed across numerous nodes;
2. the data under consideration may vary;
3. operational data can be messy, incomplete, and distributed over numerous systems

practical challenges

1. systems may have many instances, sometimes non-identical; usage may be distributed across numerous nodes;
2. the data under consideration may vary;
3. operational data can be messy, incomplete, and distributed over numerous systems

practical challenges

1. systems may have many instances, sometimes non-identical; usage may be distributed across numerous nodes;
2. the data under consideration may vary;
3. operational data can be messy, incomplete, and distributed over numerous systems

... where academic test environments are clean, simple, and ideal

3. reliability vs validity

many academic test frameworks are driven by curiosity or opportunity

customers have complex needs

which may involve combinations of information

1. “Is this political question worth taking a stand on?”
2. “What factors appear to worry potential customers for our product at what stage in their purchase path?”
3. “What factors in the pension system cause most confusion for our senior citizens?”
4. “Does this group of people pose a risk for public safety?”
5. “Will it be easy or difficult to recruit college graduates to this business area next Fall?”

customers have complex needs

which may involve combinations of information

1. “Is this political question worth taking a stand on?”
2. “What factors appear to worry potential customers for our product at what stage in their purchase path?”
3. “What factors in the pension system cause most confusion for our senior citizens?”
4. “Does this group of people pose a risk for public safety?”
5. “Will it be easy or difficult to recruit college graduates to this business area next Fall?”

customers have complex needs

which may involve combinations of information

1. “Is this political question worth taking a stand on?”
2. “What factors appear to worry potential customers for our product at what stage in their purchase path?”
3. “What factors in the pension system cause most confusion for our senior citizens?”
4. “Does this group of people pose a risk for public safety?”
5. “Will it be easy or difficult to recruit college graduates to this business area next Fall?”

customers have complex needs

which may involve combinations of information

1. “Is this political question worth taking a stand on?”
2. “What factors appear to worry potential customers for our product at what stage in their purchase path?”
3. “What factors in the pension system cause most confusion for our senior citizens?”
4. “Does this group of people pose a risk for public safety?”
5. “Will it be easy or difficult to recruit college graduates to this business area next Fall?”

customers have complex needs

which may involve combinations of information

1. “Is this political question worth taking a stand on?”
2. “What factors appear to worry potential customers for our product at what stage in their purchase path?”
3. “What factors in the pension system cause most confusion for our senior citizens?”
4. “Does this group of people pose a risk for public safety?”
5. “Will it be easy or difficult to recruit college graduates to this business area next Fall?”

customers have complex needs

which may involve combinations of information

1. “Is this political question worth taking a stand on?”
2. “What factors appear to worry potential customers for our product at what stage in their purchase path?”
3. “What factors in the pension system cause most confusion for our senior citizens?”
4. “Does this group of people pose a risk for public safety?”
5. “Will it be easy or difficult to recruit college graduates to this business area next Fall?”

special case: no news is good news

1. “What published work might be relevant to assessing the novelty of this potential patent application?”
2. “Did our customers notice that we mis-labeled the content of our product and corrected it and if they do, do they care?”

special case: no news is good news

1. “What published work might be relevant to assessing the novelty of this potential patent application?”
2. “Did our customers notice that we mis-labeled the content of our product and corrected it and if they do, do they care?”

special case: no news is good news

1. “What published work might be relevant to assessing the novelty of this potential patent application?”
2. “Did our customers notice that we mis-labeled the content of our product and corrected it and if they do, do they care?”

data are dynamic

1. “Do items posted on that video streaming site infringe on our copyright?”
2. “Is the pricing of this tradeable asset moving in some direction?”
3. “How should we set the initial odds for this bet in our book?”
4. “Will the data from our newly acquired division merge well with what we have been working on before?”
5. “What are some of the more interesting trends in our market area that are likely to influence our sales five years down the road?”

data are dynamic

1. “Do items posted on that video streaming site infringe on our copyright?”
2. “Is the pricing of this tradeable asset moving in some direction?”
3. “How should we set the initial odds for this bet in our book?”
4. “Will the data from our newly acquired division merge well with what we have been working on before?”
5. “What are some of the more interesting trends in our market area that are likely to influence our sales five years down the road?”

data are dynamic

1. “Do items posted on that video streaming site infringe on our copyright?”
2. “Is the pricing of this tradeable asset moving in some direction?”
3. “How should we set the initial odds for this bet in our book?”
4. “Will the data from our newly acquired division merge well with what we have been working on before?”
5. “What are some of the more interesting trends in our market area that are likely to influence our sales five years down the road?”

data are dynamic

1. “Do items posted on that video streaming site infringe on our copyright?”
2. “Is the pricing of this tradeable asset moving in some direction?”
3. “How should we set the initial odds for this bet in our book?”
4. “Will the data from our newly acquired division merge well with what we have been working on before?”
5. “What are some of the more interesting trends in our market area that are likely to influence our sales five years down the road?”

data are dynamic

1. “Do items posted on that video streaming site infringe on our copyright?”
2. “Is the pricing of this tradeable asset moving in some direction?”
3. “How should we set the initial odds for this bet in our book?”
4. “Will the data from our newly acquired division merge well with what we have been working on before?”
5. “What are some of the more interesting trends in our market area that are likely to influence our sales five years down the road?”

data are dynamic

1. “Do items posted on that video streaming site infringe on our copyright?”
2. “Is the pricing of this tradeable asset moving in some direction?”
3. “How should we set the initial odds for this bet in our book?”
4. “Will the data from our newly acquired division merge well with what we have been working on before?”
5. “What are some of the more interesting trends in our market area that are likely to influence our sales five years down the road?”

when the pattern is more interesting than the data points

how well a system meets such needs is difficult to evaluate

1. analysis of results may involve several steps beyond the retrieval
2. value may be impossible to assess at search time
3. what is relevant may be decided someone other than the person who formulates the information need

how well a system meets such needs is difficult to evaluate

1. analysis of results may involve several steps beyond the retrieval
2. value may be impossible to assess at search time
3. what is relevant may be decided someone other than the person who formulates the information need

how well a system meets such needs is difficult to evaluate

1. analysis of results may involve several steps beyond the retrieval
2. value may be impossible to assess at search time
3. what is relevant may be decided someone other than the person who formulates the information need

how well a system meets such needs is difficult to evaluate

1. analysis of results may involve several steps beyond the retrieval
2. value may be impossible to assess at search time
3. what is relevant may be decided someone other than the person who formulates the information need

clean canned data in a crisp task setting do not begin to approximate what is going in such cases

use cases!

if there is no use case it may be implicit

4. information pipeline

1. traditional industrial practice does not prioritise quality metrics
2. customers make purchase decisions by numerous criteria: platform, scalability, reliability, fit to other systems, maintenance; content quality is taken to be constant
3. feedback from end users is handled by customer service, not devops, thru workarounds or customer training
4. continuous improvement in devops needs information pipeline

4. information pipeline

1. traditional industrial practice does not prioritise quality metrics
2. customers make purchase decisions by numerous criteria: platform, scalability, reliability, fit to other systems, maintenance; content quality is taken to be constant
3. feedback from end users is handled by customer service, not devops, thru workarounds or customer training
4. continuous improvement in devops needs information pipeline

4. information pipeline

1. traditional industrial practice does not prioritise quality metrics
2. customers make purchase decisions by numerous criteria: platform, scalability, reliability, fit to other systems, maintenance; content quality is taken to be constant
3. feedback from end users is handled by customer service, not devops, thru workarounds or customer training
4. continuous improvement in devops needs information pipeline

4. information pipeline

1. traditional industrial practice does not prioritise quality metrics
2. customers make purchase decisions by numerous criteria: platform, scalability, reliability, fit to other systems, maintenance; content quality is taken to be constant
3. feedback from end users is handled by customer service, not devops, thru workarounds or customer training
4. continuous improvement in devops needs information pipeline

4. information pipeline

1. traditional industrial practice does not prioritise quality metrics
2. customers make purchase decisions by numerous criteria: platform, scalability, reliability, fit to other systems, maintenance; content quality is taken to be constant
3. feedback from end users is handled by customer service, not devops, thru workarounds or customer training
4. continuous improvement in devops needs information pipeline

interpreting results

1. operational systems are evaluated by sales and customer satisfaction
2. components are evaluated by engineering departments
3. there are many many components and many may unit tests
4. unit test are typically pass/fail; evaluation scores are not

interpreting results

1. operational systems are evaluated by sales and customer satisfaction
2. components are evaluated by engineering departments
3. there are many many components and many may unit tests
4. unit test are typically pass/fail; evaluation scores are not

interpreting results

1. operational systems are evaluated by sales and customer satisfaction
2. components are evaluated by engineering departments
3. there are many many components and many may unit tests
4. unit test are typically pass/fail; evaluation scores are not

interpreting results

1. operational systems are evaluated by sales and customer satisfaction
2. components are evaluated by engineering departments
3. there are many many components and many may unit tests
4. unit test are typically pass/fail; evaluation scores are not

interpreting results

1. operational systems are evaluated by sales and customer satisfaction
2. components are evaluated by engineering departments
3. there are many many components and many may unit tests
4. unit test are typically pass/fail; evaluation scores are not

interpreting results

1. operational systems are evaluated by sales and customer satisfaction
2. components are evaluated by engineering departments
3. there are many many components and many may unit tests
4. unit test are typically pass/fail; evaluation scores are not

-> engineering teams need guidance from academia sites to interpret tests

understanding evaluation results and their impact

1. public test results may have impact on
   - contractual obligations,
   - real or perceived commercial risk, or
   - user privacy
2. management may be upset when they hear of iffy results

understanding evaluation results and their impact

1. public test results may have impact on
   - contractual obligations,
   - real or perceived commercial risk, or
   - user privacy
2. management may be upset when they hear of iffy results

understanding evaluation results and their impact

1. public test results may have impact on
   - contractual obligations,
   - real or perceived commercial risk, or
   - user privacy
2. management may be upset when they hear of iffy results

understanding evaluation results and their impact

1. public test results may have impact on
   - contractual obligations,
   - real or perceived commercial risk, or
   - user privacy
2. management may be upset when they hear of iffy results

understanding evaluation results and their impact

1. public test results may have impact on
   - contractual obligations,
   - real or perceived commercial risk, or
   - user privacy
2. management may be upset when they hear of iffy results

in summary

1. evaluation must understand optimisation vs. best practice
2. evaluation schemes must be conveniently available even for commercial entities
3. evaluation metrics must have validity with respect to tasks
4. industrial organisations work towards a culture of continuous improvement
   - evaluation can be part of that culture
   - an information pipeline is needed to support such a culture
   - industrial sites will need help from academia to interpret evaluation scores
   - evaluation schemes must be doable without public results

in summary

1. evaluation must understand optimisation vs. best practice
2. evaluation schemes must be conveniently available even for commercial entities
3. evaluation metrics must have validity with respect to tasks
4. industrial organisations work towards a culture of continuous improvement
   - evaluation can be part of that culture
   - an information pipeline is needed to support such a culture
   - industrial sites will need help from academia to interpret evaluation scores
   - evaluation schemes must be doable without public results

in summary

1. evaluation must understand optimisation vs. best practice
2. evaluation schemes must be conveniently available even for commercial entities
3. evaluation metrics must have validity with respect to tasks
4. industrial organisations work towards a culture of continuous improvement
   - evaluation can be part of that culture
   - an information pipeline is needed to support such a culture
   - industrial sites will need help from academia to interpret evaluation scores
   - evaluation schemes must be doable without public results

in summary

1. evaluation must understand optimisation vs. best practice
2. evaluation schemes must be conveniently available even for commercial entities
3. evaluation metrics must have validity with respect to tasks
4. industrial organisations work towards a culture of continuous improvement
   - evaluation can be part of that culture
   - an information pipeline is needed to support such a culture
   - industrial sites will need help from academia to interpret evaluation scores
   - evaluation schemes must be doable without public results

in summary

1. evaluation must understand optimisation vs. best practice
2. evaluation schemes must be conveniently available even for commercial entities
3. evaluation metrics must have validity with respect to tasks

(validity can be achieved e.g. through use cases)

4. industrial organisations work towards a culture of continuous improvement
   - evaluation can be part of that culture
   - an information pipeline is needed to support such a culture
   - industrial sites will need help from academia to interpret evaluation scores
   - evaluation schemes must be doable without public results

in summary

1. evaluation must understand optimisation vs. best practice
2. evaluation schemes must be conveniently available even for commercial entities
3. evaluation metrics must have validity with respect to tasks

   (validity can be achieved e.g. through use cases)
4. industrial organisations work towards a culture of continuous improvement
   - evaluation can be part of that culture
   - an information pipeline is needed to support such a culture
   - industrial sites will need help from academia to interpret evaluation scores
   - evaluation schemes must be doable without public results

in summary

1. evaluation must understand optimisation vs. best practice
2. evaluation schemes must be conveniently available even for commercial entities
3. evaluation metrics must have validity with respect to tasks

   (validity can be achieved e.g. through use cases)
4. industrial organisations work towards a culture of continuous improvement
   - evaluation can be part of that culture
   - an information pipeline is needed to support such a culture
   - industrial sites will need help from academia to interpret evaluation scores
   - evaluation schemes must be doable without public results

in summary

1. evaluation must understand optimisation vs. best practice
2. evaluation schemes must be conveniently available even for commercial entities
3. evaluation metrics must have validity with respect to tasks

   (validity can be achieved e.g. through use cases)
4. industrial organisations work towards a culture of continuous improvement
   - evaluation can be part of that culture
   - an information pipeline is needed to support such a culture
   - industrial sites will need help from academia to interpret evaluation scores
   - evaluation schemes must be doable without public results

in summary

1. evaluation must understand optimisation vs. best practice
2. evaluation schemes must be conveniently available even for commercial entities
3. evaluation metrics must have validity with respect to tasks

   (validity can be achieved e.g. through use cases)
4. industrial organisations work towards a culture of continuous improvement
   - evaluation can be part of that culture
   - an information pipeline is needed to support such a culture
   - industrial sites will need help from academia to interpret evaluation scores
   - evaluation schemes must be doable without public results

in summary

1. evaluation must understand optimisation vs. best practice
2. evaluation schemes must be conveniently available even for commercial entities
3. evaluation metrics must have validity with respect to tasks

   (validity can be achieved e.g. through use cases)
4. industrial organisations work towards a culture of continuous improvement
   - evaluation can be part of that culture
   - an information pipeline is needed to support such a culture
   - industrial sites will need help from academia to interpret evaluation scores
   - evaluation schemes must be doable without public results