QUADS: Question Answering for Decision Support

Zi Yang¹, Ying Li², James Cai², Eric Nyberg¹

¹) Carnegie Mellon University   {ziy, ehn}@cs.cmu.edu
²) Roche Innovation Center   {ying_l.li, james.cai}@roche.com

09/07/2014 at SIGIR 2014
Decision Making

*Everyday life decision making*

Which phone (📱,📲,📞) is recommended to purchase?

*Scientific decision making*

Which disease target (균, 독, 바이러스) should I explore the potential for making a drug?
Decision Analysis Theories

• People are
  – **NOT GOOD** at making high-level decisions.
    • Which phone is recommended to purchase?
  – **GOOD** at making low-level / atomic decisions.
    • Is this phone running Android?

• Analytic thinking: decompose decision factors iteratively!

Two examples.
Decision Making: Product Recommendation

**Decision decomposition**

- Design and usability
- Brand
- Functionality
- Carrier
- Operating system

- Weight
- Thickness
- Resolution
- Keyboard

**Evidence gathering from Web**

**Synthesis**

<table>
<thead>
<tr>
<th>Brand</th>
<th>Carrier</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>aaa</td>
<td>xxx</td>
<td>Good</td>
</tr>
<tr>
<td>bbb</td>
<td>yyy</td>
<td>OK</td>
</tr>
<tr>
<td>ccc</td>
<td>zzz</td>
<td>Bad</td>
</tr>
</tbody>
</table>
Decision Making: Target Validation

• Disease target
  – An agent with a particular biological action that are anticipated to have therapeutic utility.

• Target validation
  – Select and prioritize a number of disease targets and estimate the “druggability” of each target.

• Drug development pipeline
Decision Making: Target Validation

**Decision decomposition**

- Modulating the activity
- Expression in tissues
- Mutation
- Clinical trials
- Side effects

**Evidence gathering from public/proprietary documents**

- In vivo
- In vitro
- Normal tissues
- Disease tissues

**Synthesis**

<table>
<thead>
<tr>
<th>In vivo</th>
<th>Side effect</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>No</td>
<td>Good</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>OK</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>Bad</td>
</tr>
</tbody>
</table>

Answer of “no” is desired!
Decision Support Systems

• How can we automate the manual process?
  – Decision support systems (DSS), business intelligent (BI) systems, decision engines, etc.

• Can they process arbitrary *ad hoc* decision problem described in natural language?
  – *No!* They are designed for specific tasks.
  – QUADS *(Question Answering for Decision Support)*

Does modulating the activity of the GENE cause the disease?

Modulating the activity
QUADS

Known Decisions with Decision Processes (Training)

Question Answering Pipeline

Decision Process (Test)

Decision Strategy Learning

Decision Making & Reporting

Decisions Augmented with Evidence & Reasoning

QA Based Decision Factor Estimation Modeling

Decision Factor Synthesis Modeling

Decision Making & Reporting
Decision problems described in natural language

Question answering technology to automatically process the decisions

Bayesian model to infer the optimal decision.

Learning decision strategies
Outline

• Introduction
• Decision Making in QA Context
• QUADS Framework
• QUADS for Target Validation
• QUADS for Product Recommendation
• Related Work
• Conclusion & Future Work
Outline

• Introduction
• Decision Making in QA Context
• QUADS Framework
• QUADS for Target Validation
• QUADS for Product Recommendation
• Related Work
• Conclusion & Future Work
Decision Factor, Dependency

**Decision goal** $f_0$ with alternatives / candidates $a = \{a_k\}_k$

**Recommended phone?**

Which OS?

Which carrier?

**Latent variable for true supportiveness of each factor (outcomes $a$)**

**Decision factor** $f_i$

**Primitive factors** ($F_i = \emptyset$)

all dependent factors

Reflect personal preferences or prior knowledge

- Specify $p(y_i)$ to prioritize the options to your preference
- Specify conditional dependency $p(y_i|y_j)$ to weight the factors
Decision Process, QA

**Decision process (DP)**

\[ S = (Y, E) \]

- Decision factors
- Conditional dependencies

**QA produces assertions.**

**Result** estimated by the QA process

\[ o = \{o_{ik}\}_k \]

- Outcomes of yes/no questions: \{YES, NO\}.
- Outcomes of factoid questions: all candidate answers, e.g. what is the brand of phone \(a_k\): all phone brands.

**Benefit of separation of X and Y:**
- Different outcomes: \(o\) vs \(a\)
- Estimated vs true supportiveness
- Capture positive/negative support
Decision Making Problem

\[
(\text{Decision Process}, \text{Question Answering Pipeline}) \rightarrow \max_{k} \text{Utility}(a_k)
\]
Outline

• Introduction
• Decision Making in QA Context
• QUADS Framework
• QUADS for Target Validation
• QUADS for Product Recommendation
• Related Work
• Conclusion & Future Work
**Assumptions:**

1. QA is done individually, i.e. $x_i \perp x_j$.
2. Variable $x_i$ is conditioned on $y_i$ following \textit{estimation policy} $\xi_i: a \rightarrow o_i$.

**Decision synthesized from** $f_i$ and $F_i$.

**Assumptions:**

1. Only depends on immediate predecessors, i.e. $d_i | F_i$.
2. The outcomes by $d_i$ are identical to $o_i$, following \textit{synthesis policy} $\delta_i: a \times \prod_{k \in F_i} o_k \rightarrow o_i$.
Modeling Decision Process

• Joint probability to model QUADS framework

\[ P(Y, D, X; \Delta, \Xi) = \prod_i p(y_i | y_{Fi}) p^\xi(x_i | y_i) p^\delta(d_i | y_i, d_{Fi}) \]

- Decision maker’s preference
- QA based decision factor estimation components
- Factor synthesis components
Solution

• Decision Making
  – Finding the values of Y and D that maximize the conditional probability
    \[ Y^*, D^* = \arg \max_{Y,D} P(Y, D|X; \Delta, \Xi) \]
    – Optimal solution can be obtained efficiently if each decision factor only belongs to one super-factor, i.e.
      \[ F_i \cap F_j = \emptyset. \]

• Learning decision strategies
  – The policies \( \Delta \) and \( \Xi \) are unknown
  – Since utility function is degenerated, the local MEU problem becomes MLE.
QUADS Implementation

• Compatible with UIMA and CSE Framework
  – Integrate existing UIMA based QA pipelines (*CPE descriptors*) and decision synthesis components for each factor at each level.
  – Allow multiple QA pipelines or factor synthesis components to be specified as alternatives for the same decision factor.

• Decision process template (DPT)
  – Define variables (*GENE, DISEASE, PHONE*) at front!
  – Assign values later.
  – Shared the same policy.
Outline

• Introduction
• Decision Making in QA Context
• QUADS Framework
• **QUADS for Target Validation**
• QUADS for Product Recommendation
• Related Work
• Conclusion & Future Work
Preparation: Data Set

• Corpus for evidence supporting
  – 22M PubMED abstracts
  – Minimal preprocessing: tokenization, downcasing, removing common English stopwords
  – Indexed by Lucene

• Known gene-disease relation DBs
  – OMIM, UniProt, PharmGKB, CTD, DisGeNET
  – 5,605 diseases with 10-fold cross validation
  – 6,158 genes as candidates for the DP
Preparation: Target Validation DPT

• Based on experts’ knowledge and experience, manually created 16 factors in a 3-level hierarchy.
• Yes/No questions at each node for simplicity.
  – QA pipelines are only required to return either yes or no.
• We define **GENE** and **DISEASE** as the variables.

```
1 the GENE is directly involved in the DISEASE and can be a suitable target.
  1.1 Any experiment showing that modulating the activity of the GENE with a chemical compound or genetic modification causes the DISEASE.
  1.1.1 Any human in vivo experiment showing that modulating the activity of the GENE affects biochemical function or phenotype of the DISEASE.
  1.1.2 Any in vitro experiment showing that modulating the activity of the GENE affects biochemical function or phenotype of the DISEASE.
  1.1.3 Any animal model study showing that modulating the activity of the GENE in animals causes the DISEASE.
  1.2 the GENE is expressed in the human tissue related to the DISEASE.
    1.2.1. the GENE is expressed in normal human tissue related the the DISEASE.
    1.2.2 the GENE expression is altered in human DISEASE tissue or human DISEASE cell.
    1.2.3 The alteration of the GENE expression is correlated with the DISEASE severity.
  1.3 Any mutation is associated with the DISEASE.
    1.3.1 Any mutation of the GENE has significantly associated with the DISEASE.
    1.3.2 Any mutations in other genes linked to the GENE associated with DISEASE.
  1.4 Any pathway involving the GENE supports that the GENE causes the DISEASE.
  1.5 Any clinical trials show that targeting the GENE can prevent or slow the progress of the DISEASE.
  1.6 Any evidence suggests that targeting the GENE will have side effects.
    1.6.1 Targeting the GENE will cause liver, heart, and kidney damage.
```

“Side effect” related factors
Preparation: Supporting evidence based QA pipeline

<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>COMPONENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parsing</td>
<td>LingPipe HMM based tokenizer</td>
</tr>
<tr>
<td></td>
<td>LingPipe HMM based POS tagger</td>
</tr>
<tr>
<td></td>
<td>LingPipe HMM based named entity recognizer</td>
</tr>
<tr>
<td></td>
<td>Rule based lexical variant generator</td>
</tr>
<tr>
<td>Interpretation</td>
<td>UMLS for disease syn/acronym expansion</td>
</tr>
<tr>
<td></td>
<td>EntrezGene for gene syn/acronym expansion</td>
</tr>
<tr>
<td></td>
<td>OMIM for disease syn/acronym expansion</td>
</tr>
<tr>
<td>Retrieval</td>
<td>Lucene search engine</td>
</tr>
<tr>
<td>Evidencing &amp; ranking</td>
<td>Document frequency</td>
</tr>
<tr>
<td></td>
<td>Weight learning with logistic regression</td>
</tr>
<tr>
<td>CATEGORY</td>
<td>COMPONENTS</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Parsing</td>
<td>LingPipe HMM based tokenizer</td>
</tr>
<tr>
<td></td>
<td>LingPipe HMM based POS tagger</td>
</tr>
<tr>
<td></td>
<td>LingPipe HMM based named entity recognizer</td>
</tr>
<tr>
<td></td>
<td>Rule based lexical variant generator</td>
</tr>
<tr>
<td>Interpretation</td>
<td>UMLS for disease syn/acronym expansion</td>
</tr>
<tr>
<td></td>
<td>EntrezGene for gene syn/acronym expansion</td>
</tr>
<tr>
<td></td>
<td>OMIM for disease syn/acronym expansion</td>
</tr>
<tr>
<td>Retrieval</td>
<td>Lucene search engine</td>
</tr>
<tr>
<td>Evidencing &amp; ranking</td>
<td>Document frequency</td>
</tr>
<tr>
<td></td>
<td>Weight learning with logistic regression</td>
</tr>
</tbody>
</table>

Queries constructed from different subsets of the focus terms, nouns, verbs, NEs, syn/acronyms with **OR** to obtain a series of DF scores.

Binary classifier is trained based on FVs of DF scores.
Preparation: Evaluation Methods

• QUADS output vs gold standard

<table>
<thead>
<tr>
<th>QUADS OUTPUT</th>
<th>GOLD STANDARD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ordered lists of genes prioritized by the confidence for each candidate</td>
<td>Unordered known target lists from DisGeNET.</td>
</tr>
</tbody>
</table>

• Evaluation methods

<table>
<thead>
<tr>
<th>OUTPUT ORDER IGNORED</th>
<th>ORDER CONSIDERED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision, recall, F-1 averaged over all subsets</td>
<td>MAP</td>
</tr>
</tbody>
</table>
Preparation: Baseline Methods

• Vary DPT:
  – 1 Level, 2 Levels: leave the first one/two levels
  – Flattened: keep the factors, ignore the hierarchy

• Vary synthesis algorithm
  – Voting: choose the majority decision from all factors at each level
  – Simplified: remove latent variable $Y$

• Vary QA
  – No Synonym: no acr/synonym expansion for diseases/genes
Experimental Results: Comparison with Baseline Methods

![Graph showing comparison of experimental results with baseline methods for Precision, Recall, F-1, and MAP across different methods: QUADS, 1 LEVEL, 2 LEVEL, FLATTEN, VOTING, SIMPLIFIED, and NO SYNONYM.](image)
Experimental Results: Comparison with Baseline Methods

QUADS significantly outperformed all baseline methods in all metrics.

MAP > Precision ➔ Most extensively studied genes are more likely to be selected correctly.

High Recall, low Precision ➔ Potentially correct outputs might be missing in the GS set.
Experimental Results: Comparison with Baseline Methods

0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9

Precision  Recall  F-1  MAP

QUADS  1 LEVEL  2 LEVEL  FLATTEN  VOTING  SIMPLIFIED  NO SYNONYM
Experimental Results: Comparison with Baseline Methods

1 Level & 2 Level: low performance due to low recall, missing related information

Flattened & Voting: lower MAP than P/R, cannot synthesize scores. In fact, only OR / weighted SUM is used for policy $\delta$.

Simplified: performance very close to QUADS.
No Synonym: the worst!
Experimental Analysis: Learned Estimation Policy

\[ p_i^\xi (x_i = \text{yes}|y_i = \text{yes}) \quad p_i^\xi (x_i = \text{no}|y_i = \text{no}) \]
Experimental Analysis: Learned Estimation Policy

The most general question (decision goal) will be the hardest for QA to answer.

The policy for a perfect QA system should have diagonal values as 1 if the factors can positively contribute, i.e. 1 for both.

Or 0 if negatively contribute, e.g. 1.6 and 1.6.1

The learned policies are consistent with out intuitions.
Experimental Analysis: Top-ranked Decision Policies

| $f$ | Rank | $y_i, d_{Fi}$ | $p_i^\delta(d_i = yes|y_i, d_{Fi})$ |
|-----|------|---------------|----------------------------------|
| 1.1 | 1    | $y_{1.1} = yes, d_{1.1.*} = yes$ | .9984 |
|     | 2    | $y_{1.1} = no, d_{1.1.*} = yes$ | .9561 |
| 1.3 | 1    | $y_{1.3} = yes, d_{1.3.*} = yes$ | .9957 |
|     | 2    | $y_{1.3} = no, d_{1.3.*} = yes$ | .9451 |
| 1   | 1    | $y_1 = yes, d_{1.*} = yes$ | .9710 |
|     | 2    | $y_1 = no, d_{1.*} = yes$ | .5015 |
Experimental Analysis: Top-ranked Decision Policies

| $f$ | Rank | $y_i, d_{F_i}$          | $p_i^\delta(d_i = \text{yes} | y_i, d_{F_i})$ |
|-----|------|-------------------------|-----------------|
| 1.1 | 1    | $y_{1.1} = \text{yes}, d_{1.1.\ast} = \text{yes}$ | .9984           |
|     | 2    | $y_{1.1} = \text{no}, d_{1.1.\ast} = \text{yes}$  | .9561           |
| 1.3 | 1    | $y_{1.3} = \text{yes}, d_{1.3.\ast} = \text{yes}$ | .9957           |
|     | 2    | $y_{1.3} = \text{no}, d_{1.3.\ast} = \text{yes}$  | .9451           |
| 1   | 1    | $y_1 = \text{yes}, d_{1.\ast} = \text{yes}$       | .9710           |
|     | 2    | $y_1 = \text{no}, d_{1.\ast} = \text{yes}$        | .5015           |

Always prefer all “yes”

$f_1$ is harder than others

The local estimation $y$ can be ignored.
Breast Cancer: Qualitative Judgment

• Recall: 100%

• Manual qualitative judgment
  – Qualified targets: AKT1, BRCA1, BRCA2, BRIP1, ERBB2, FGFR1, KDR1, mTOR, and NBN
  – High confidence scores given by QUADS
  – Not included in DisGeNET: FGFR1, KDR1, mTOR
Breast Cancer: Quantitative Analysis

# of OR-combined query terms

(a) TF

(b) FP

(c) TN

DF
Breast Cancer: Quantitative Analysis

More evidence gathered in TF than FP and TN

Answerability of questions differs across factors

# of OR-combined query terms

DF

(a) TF

(b) FP

(c) TN
Outline

• Introduction
• Decision Making in QA Context
• QUADS Framework
• QUADS for Target Validation
• QUADS for Product Recommendation
• Related Work
• Conclusion & Future Work
Preparation: Dataset

• Goal
  – Adapt the QUADS framework to product recommendation with minimal change to test the flexibility and generalizability.

• Corpus
  – 78,931 review texts for 7,438 products related to Phone and Accessories from Amazon product review.
  – 5-point ratings were averaged and converted to binary recommendation. (yes if 4 or 5, no o.w.)
  – 10-fold cross validation over products.
Preparation: Decision Process Template

• Decision goal
  – Buying this PHONE is recommended?

• Factors
  – Design and usability, brand, functionality, carrier, and operating system, etc.
  – Allow QA pipelines to return nominal answer texts.
  – 3-level hierarchy with 17 factors.
Preparation: QA Pipeline

- Adapt almost the same components
- Changes
  - Replace POS tagging and NER model with models trained on English news corpus.
  - Replace biomedical synonym expansion with a sentiment word dictionary.
  - Integrate the review text index to replace the PubMed abstract index.
Experimental Results & Analysis

• Compare with baseline methods
Experimental Results & Analysis

- Compare with baseline methods

Lower scores ➔ Everyday decision is harder to be made than scientific decision.

Target validation
Navigating through millions of publications

Product recommendation
Informal text and mixed reviews
Outline

• Introduction
• Decision Making in QA Context
• QUADS Framework
• QUADS for Target Validation
• QUADS for Product Recommendation
• Related Work
• Conclusion & Future Work
Related Work

• Decision analysis models
  – AHP [Saaty, 1988], ID [Howard & Matheson, 1987]

• Decision support systems based on natural language input
  – Structured Evidential Argumentation System (SEAS) [Knowles & Gromo, 2003]
  – Intelligent Decision System (IDS) [Xu et al., 2006]

• Question decomposition for factoid QA

• Scenario & contextual QA
Outline

• Introduction
• Decision Making in QA Context
• QUADS Framework
• QUADS for Target Validation
• QUADS for Product Recommendation
• Related Work
• Conclusion & Future Work
Conclusion

• A novel decision representation based on natural language questions or assertions.
• QUADS framework and implementation that takes advantage of QA technologies.
• Case studies on two different problems: target validation and product recommendation from text
Future Work

• Improve QA pipelines, such as adding structured information retrieval.

• A joint optimization of question answering and decision making becomes possible.

• Automatic decision process construction.

• Allow the decision process to emerge dynamically, with selective pruning of alternatives.
Thanks!
Now I’m glad to take your questions.