2nd Idea

Visualize with graph: Easy to see relationship to other questions
Motivation: online visual word dictionary

Steps 1: Massage the data using script

1. Understand the data attribute
2. Determine what field will be used for what purpose
3. Write script to generate massaged dataset

Step 2: Analyze the data using R

Three criteria of questions

1. Size = quality
   << Posts.xml >>
   a. Score
   ~~b. View Count~~ i.e) after normalization the max is still over thousand...
   c. Answer Count
   d. Comment Count
   e. Favorite Count
2. distance = similarity [content based]
   << Posts.xml --> Body>>
   Ignored Title; too short. not always similar. the content is more important
   a. Apply Vector Space Model based on keywords in questions
      i. exclude stop words
      ii. go around or solve spelling errors, typos
      iii. synonyms (search engine)
   b. Use Cosine Distance
   c. ~~Frequency count~~ or boolean? i.e) doesn't matter how many times appears
3. thickness = association
   a. Postlink: one question or answer link to another question or answer
      Analyze postlink.xml; some post in PostId is deleted --> can find out when applying to the post.xml
   b. Comments: PostId ref to Id in posts.xml
      Text: … http://askubuntu.com/questions/*78352*/... --> link to question
      http://askubuntu.com/users/1992/roland-taylor → link to user
   c. ~~Posts → ownerUserId: Two questions are asked by the same user or answers answered by the same user~~ *c.f) Need to combine question and its answers*

Step 3: Visualize the result using graph tool

- http://sigmajs.org/
- http://neo4j.com/
- http://gephi.github.io/
- http://www.cytoscape.org/

After making the graph, the distance between all nodes will be there

Then, use k-means or any clustering algo to cluster the questions

~~1st Idea~~

Input: some question
Output: I am 87% sure this questions relates to your question and its answer has 89% quality

Steps:

1. Extract keywords from the input i.e) stoplist
2. Find the most relevant cluster
3. With the cluster, find the most relevant question by Vector Space Model
4. Calculate score for relevance
5. Calculate score for the quality of answer

Extra

6 degree of separation by analyzing the linking among questions
Make it dynamic i.e) stream the question and answer

Detect the question which are not appropriate for this forum. Assist admin to close it.