

What is claimed is:

1. A method of generating Frequently Asked Questions (FAQ) data from Community-based Question Answering (CQA) data, the method comprising:

receiving a plurality of data sources and a topic having one or more

themes, wherein each data source has data associated with one or more topics;

generating a thematic hierarchy of the plurality of data sources;

classifying a plurality of CQA data into one or more themes based on the thematic hierarchy, where the CQA data containing a plurality of question-answer pairs;

selecting a plurality of question-answer pairs from the CQA data based on

the classification, the selecting comprising:

for each theme of the CQA data, grouping a plurality of CQA data

into a plurality of clusters, wherein the CQA data in a cluster

share one or more features associated with the theme, and a

cluster of CQA data has a centroid representing the theme of

the cluster; and

generating FQA FAQ data using the selected question-answer pairs of the CQA data.

2. The method of claim 1, wherein the topics and themes of the plurality of data sources are organized hierarchically within the thematic hierarchy.

3. The method of claim 1, wherein classifying the plurality of CQA data comprises using a centroid-based classifier, wherein a theme of the CQA data has a centroid of a plurality of prototypes associated with the theme.

4. The method of claim 3, wherein the centroid of the plurality of prototypes associated with the theme is based on the weights assigned to the plurality of prototypes associated with the theme.

~~5. The method of claim 1, wherein selecting the plurality of question-answer pairs from the CQA data comprises:~~

~~for each theme of the CQA data, grouping a plurality of CQA data into a plurality of clusters, wherein the CQA data in a cluster share one or more features associated with the theme, and a cluster of CQA data has a centroid representing the theme of the cluster.~~

6.5. The method of claim 5 1, wherein selecting the plurality of question-answer pairs from the CQA data further comprises, for each cluster of CQA data:

selecting a plurality of representative data from the cluster;
measuring a quality of the representative data; and
generating a representative score for each question-answer pairs of the representative data.

7.6. The method of claim 6 5, wherein measuring the quality of the representative data comprises generating a quality score of the representative data.

~~8.7.~~ The method of claim 6 5, wherein generating a representative score of the representative data comprises calculating the distance between a question-answer pair of the CQA data with the centroid of the cluster.

~~9.8.~~ The method of claim 6 5, wherein generating a representative score for each question-answer pair of the representative data further comprises ranking the question-answer pairs of the CQA data in the cluster based on the representative scores.

~~10.9.~~ A non-transitory computer-readable medium storing executable computer program code for generating Frequently Asked Questions (FAQ) data from Community-based Question Answering (CQA) data, the computer program code comprising code for:

receiving a plurality of data sources, a data source having data associated

with one or more topics, and a topic having one or more themes;

generating a thematic hierarchy of the plurality of data sources;

classifying a plurality of CQA data into one or more themes based on the thematic hierarchy, where the CQA data containing a plurality of question-answer pairs;

selecting a plurality of question-answer pairs from the CQA data based on

the classification, the selecting comprising:

for each theme of the CQA data, grouping a plurality of CQA data into a plurality of clusters, wherein the CQA data in a cluster share one or more features associated with the theme, and a

cluster of CQA data has a centroid representing the theme of the cluster; and
generating ~~FQA~~ FAQ data using the selected question-answer pairs of the CQA data.

~~11.10.~~ The computer-readable medium of claim ~~10~~ 9, wherein the topics and themes of the plurality of data sources are organized hierarchically within the thematic hierarchy.

~~12.11.~~ The computer-readable medium of claim ~~10~~ 9, wherein the computer program code for classifying the plurality of CQA data comprises computer program code for using a centroid-based classifier, wherein a theme of the CQA data has a centroid of a plurality of prototypes associated with the theme.

~~13. The computer-readable medium of claim 10, wherein the computer program code for selecting the plurality of question-answer pairs from the CQA data comprises computer program code for:~~

~~for each theme of the CQA data, grouping a plurality of CQA data into a plurality of clusters, wherein the CQA data in a cluster share one or more features associated with the theme, and a cluster of CQA data has a centroid representing the theme of the cluster.~~

~~14.12.~~ The computer-readable medium of claim ~~13~~ 9, wherein the computer program code for selecting the plurality of question-answer pairs from the CQA data further comprises computer program code for, for each cluster of CQA data:

selecting a plurality of representative data from the cluster;
measuring quality of the representative data; and
generating a representative score for each question-answer pairs of the
representative data.

~~15.13.~~ The computer-readable medium of claim ~~14~~ 12, wherein the computer program code for measuring the quality of the representative data comprises computer program code for generating a quality score of the representative data.

~~16.14.~~ The computer-readable medium of claim ~~14~~ 12, wherein the computer program code for generating a representative score of the representative data comprises computer program code for calculating the distance between a question-answer pair of the CQA data with the centroid of the cluster.

~~17.15.~~ The computer-readable medium of claim ~~14~~ 12, wherein the computer program code for generating a representative score for each question-answer pair of the representative data further comprises computer program code for ranking the question-answer pairs of the CQA data in the cluster based on the representative scores.

~~18.16.~~ A system of generating Frequently Asked Questions (FAQ) data from Community-based Question Answering (CQA) data, the system comprising:

a non-transitory computer-readable storage medium storing executable
computer program modules comprising a thematic hierarchy
generation module configured to:

receive a plurality of data sources, a data source having data associated with one or more topics, and a topic having one or more themes, and
generate a thematic hierarchy of the plurality of data sources;
a feature classifier configured to classify a plurality of CQA data into one or more themes based on the thematic hierarchy, where the CQA data containing a plurality of question-answer pairs; and
a selection configured to:
select a plurality of question-answer pairs from the CQA data based on the

classification, the selecting comprising:

for each theme of the CQA data, grouping a plurality of CQA data into a plurality of clusters, wherein the CQA data in a cluster share one or more features associated with the theme, and a cluster of CQA data has a centroid representing the theme of the cluster, and

generate FQA FAQ data using the selected question-answer pairs of the CQA data.

19. ~~The system of claim 18, wherein the selection module is further configured to:~~

~~for each theme of the CQA data, group a plurality of CQA data into a plurality of clusters, wherein the CQA data in a cluster share one or more features associated with the theme, and a cluster of CQA data has a centroid representing the theme of the cluster.~~

~~20~~17. The system of claim ~~19~~ 16, wherein the selection module is further configured to, for each cluster of CQA data:

select a plurality of representative data from the cluster;

measure quality of the representative data; and

generate a representative score for each question-answer pairs of the representative data.