

Mining Duplicate Questions in Stack Overflow

Muhammad Ahasanuzzaman

Muhammad Asaduzzaman

Chanchal K. Roy

Kevin A. Schneider

背景

- StackOverflowのようなCQAサイトに注目が集まっている
- StackOverflowの質問と回答には投票機能があり，投票数は大まかな質として扱われている
- ユーザーは投票数により，バッジや評価やバッジを手に入れることができる
 - バッジや評価によってStackOverflowの運営に関する特権が与えられる

課題

- 重複質問に関する課題
 - StackOverflowでは、新しい質問をする前に重複した質問を検索することを推奨している
 - しかし、重複した質問が多く投稿されているのが現状
- ユーザーへの負担
 - 重複した質問の処理は、StackOverflowのモデレータ（特別な権限を持っているユーザー）と一定数以上の評価を持っているユーザーが行なっている
 - ・ Duplicateフラグをつけて、質問をcloseにする

目的・リサーチクエスチョン

- 目的

- 重複質問の処理を行うユーザーの負担を軽減するために、
重複質問を自動特定するアプローチを提案する

- リサーチクエスチョン

- 1.なぜ重複した質問が投稿されるのか？
- 2.機械学習を用いて、重複した質問を特定する手法を実現できるか？
- 3.提案手法の精度は、既存の手法と比較してどの程度か？

データセット

- MSR2015 Mining Challengeのデータセット
 - 2008年8月~2014年9月
 - 重複した質問：130,888件
 - master質問：90,245件
- 重複質問の抽出
 - 一部のユーザーが重複した質問に「Duplicate」フラグをつけることができるため、フラグの付いている質問を抽出
 - ・ 評判が50以上のユーザーのみDuplicateフラグをつけることができる
 - 重複質問のものの質問（master）を特定

重複した質問の分析

日付と重複質問

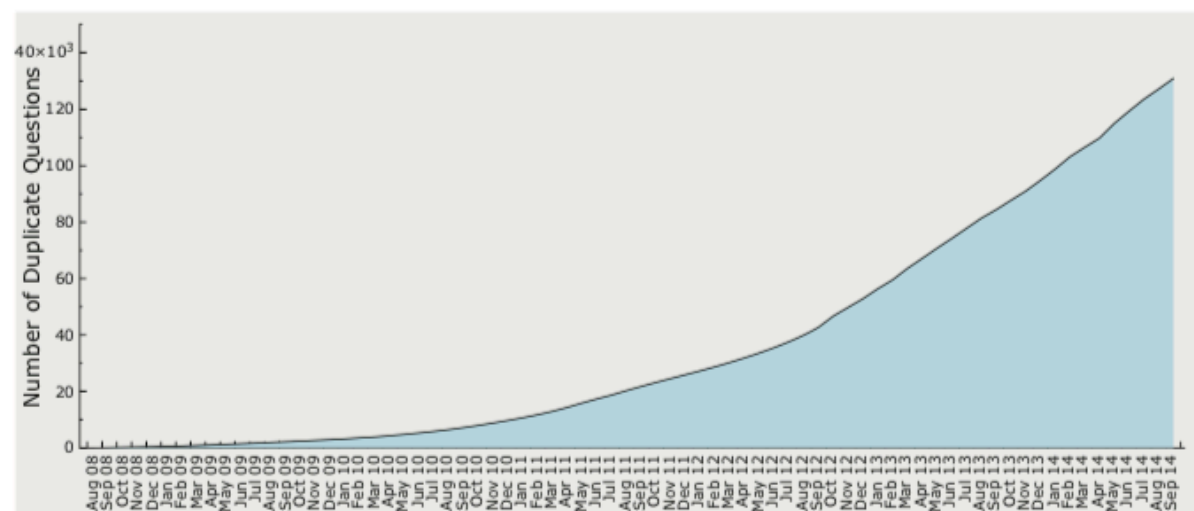


Figure 2: Cumulative distribution of duplicate questions (Aug 2008-Sep 2014). There is a sharp rise in duplicate questions after Sep 2012

重複している質問につけられているタグ

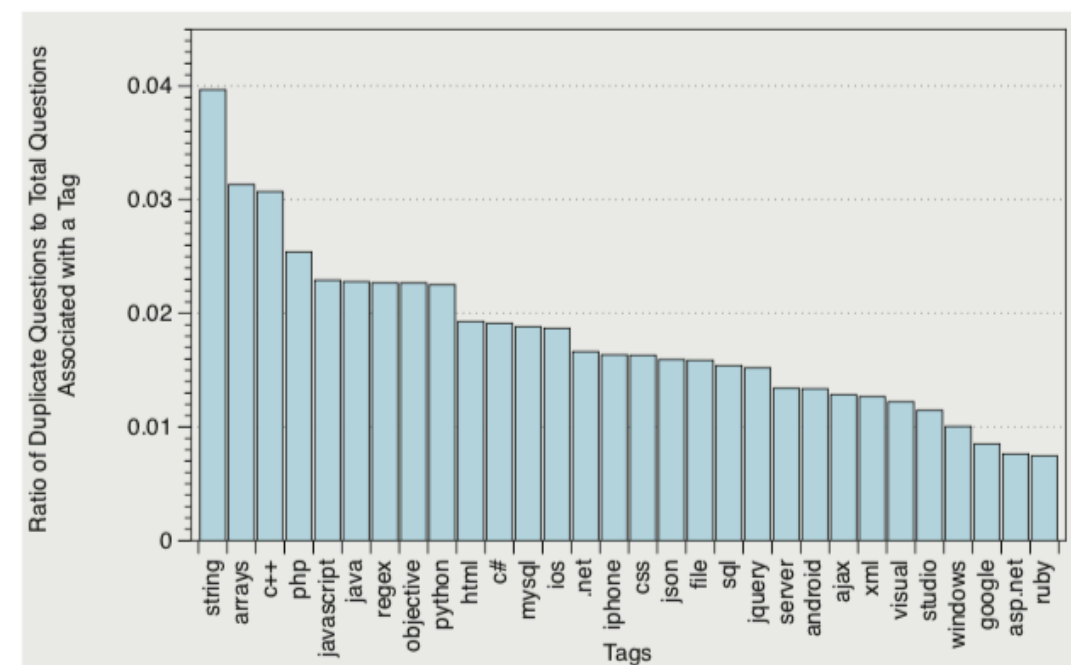


Figure 3: Distribution of duplicate questions for the most popular tags

重複した質問がcloseになるまでの日数

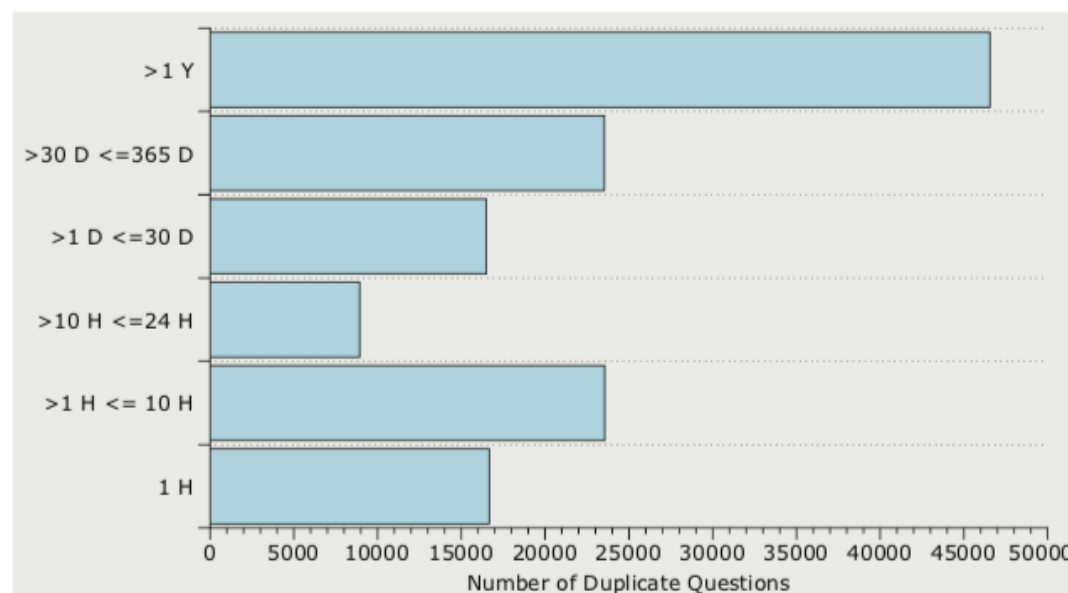


Figure 4: Time required to close duplicate questions

RQ1に関する調査

- なぜ重複した質問が投稿されるのか？
- 調査方法
 - 2009年から2014年の間から毎年100件ずつ無作為に抽出し，計600件を手動で調査する
 - ・ 重複した質問と重複元の質問を比較し，著者2人が重複した理由を議論した

ドメインが異なる事例

- どちらの質問もJavaでタイマーを作成したいという質問

How to set a timer in Java?
How to set a Timer say for 2 minutes to try to connect to a database and then exception out if there is any issue in connecting
データベースの接続をタイムアウトするための

Animation using a timer?
I'm trying to understand how to use one some kind of basic animation. I can use threads, but I was told by multiple people not to multi-thread in java. I think I cold do something like:
<pre>Timer t = new Timer(10, something); t.start(); x++; t.end();</pre>
That's my basic understanding of it. Can someone link a tutorial or explain how to make a time do something every 10 seconds or longer?
タイマーをしようしてアニメーションを作成したい

質問の内容が簡素すぎる事例

IOException: The process cannot access the file 'file path' because it is being used by another process

I have some code and when it executes, it throws a `IOException`, saying "The process cannot access the file 'filename' because it is being used by another process".

What does this mean, and what can I do about it?

Of course this is not a real question; I found many questions here on SO are related to this error so a general guidance (like this one for [NullReferenceException](#) and [IndexOutOfRangeException](#)) may be useful.

Answer is a community wiki so feel free to edit and improve it with more common scenarios and suggestions.

c# .net language-agnostic ioexception

M

Error in my code when delete file?

I edit pic file in path and create new image file for there.my code is:

```
string[] files = Directory.GetFiles(string.Concat(Server.MapPath("/"), "tmp/"));
foreach (string path in files)
{
    string filename = Path.GetFileName(path);
    using (Bitmap b = new Bitmap(string.Concat(Server.MapPath("/"), "tmp/", filename)))
    {
        SolidBrush pixelBrush = new SolidBrush(Color.White);
        Graphics g = Graphics.FromImage(b);
        g.FillRectangle(Brushes.White, 0, 0, 105, 40);

        string outputFileName = string.Concat(Server.MapPath("/"), "tmp\\E", filename);
        MemoryStream memory = new MemoryStream();
        FileStream fs = new FileStream(outputFileName, FileMode.Create, FileAccess.ReadWrite);
        b.Save(memory, ImageFormat.Jpeg);
        byte[] bytes = memory.ToArray();
        fs.Write(bytes, 0, bytes.Length);
        fs.Close();
        memory.Close();
        b.Dispose();
    }
    File.Delete(path);
}
```

when delete old file error happend is:

Additional information: The process cannot access the file
'G:\project\WebApplication1\WebApplication1\tmp\b381ae6.jpg' because it is being used by
another process.

how to fix it?

D

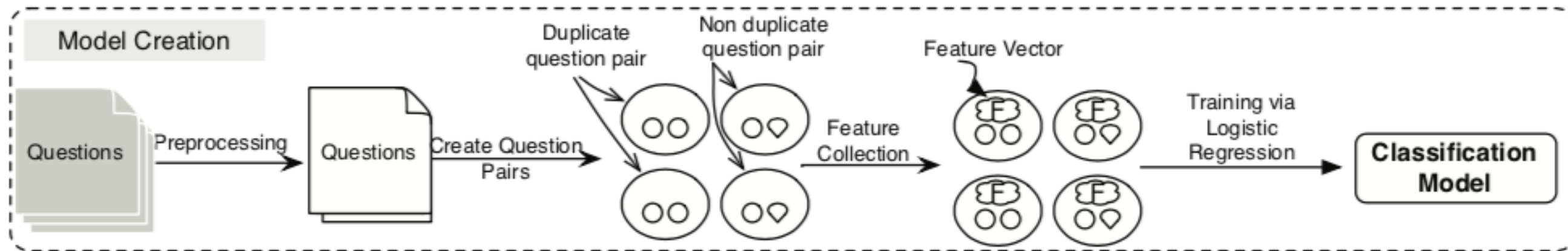
RQ1の調査結果

Table 2: Reasons behind duplicate questions

Reasons	Percent
Not searching Stack Overflow first	35.5
Lack of knowledge about the problem	19.0
Titles do not match	17.0
Too concise to properly understand	12.0
Domain difference despite task similarity	7.0
Lack knowledge of terminology or buzzwords	5.5
Descriptive and difficult to comprehend	4.0

- 質問前に検索しない，問題に関する知識が不足しているため重複しているか分からないなどの理由が多い
- 重複特定する手法を提案することで，質問の質問者とモデレータの時間を節約することができる

提案手法



- Preprocessing

- ストップワードの除去, ステミング処理
- 質問についているタグを統一
 - Tag synonymを用いて類似タグを統一

提案手法の概要

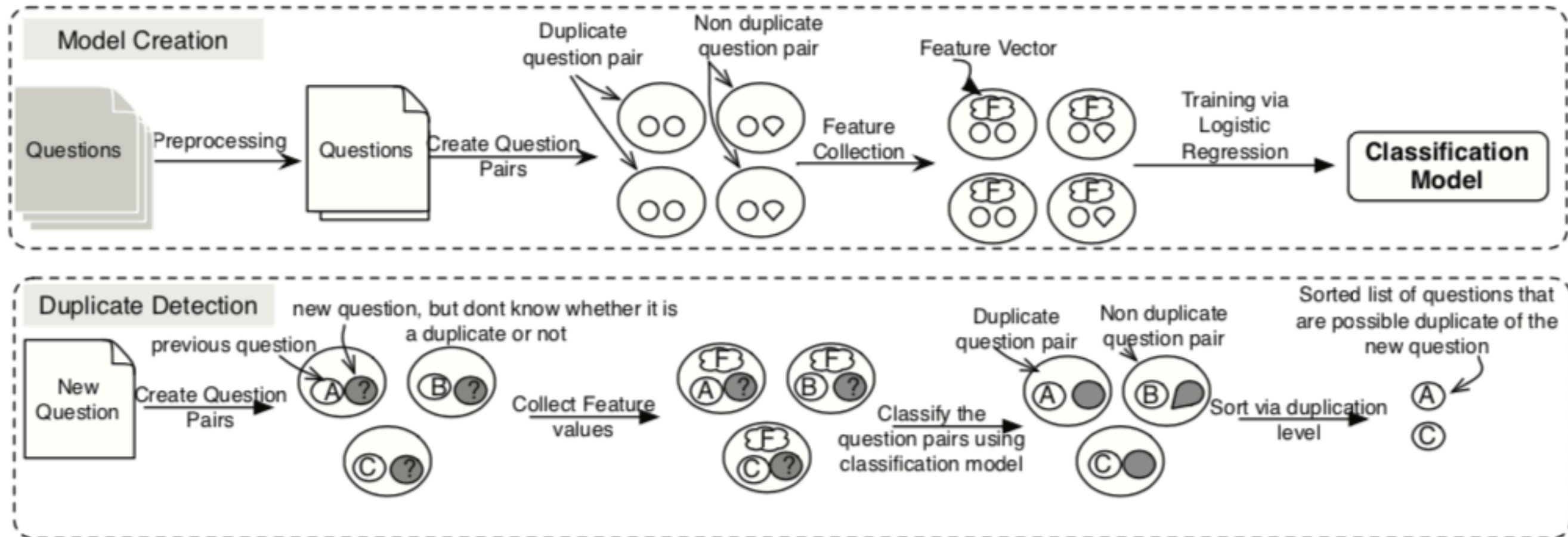


Figure 10: Classification model creation and duplicate detection via the model

特徴量

- 文書間のコサイン類似度
- 用語の重複 (t1 文書1のタイトル t2文書2のタイトル)

$$\frac{2|t_1 \cap t_2|}{|t_1| \cup |t_2|}$$

- Named Entity Recognition (NER) の重複
 - 固有名詞や人名など (Stanford Named Entity Recognizer により抽出)
- NERのタイプの重複
- WordNet Similarity
 - WordNetの辞書を利用して, 文書の類似性を測る

モデルの作成

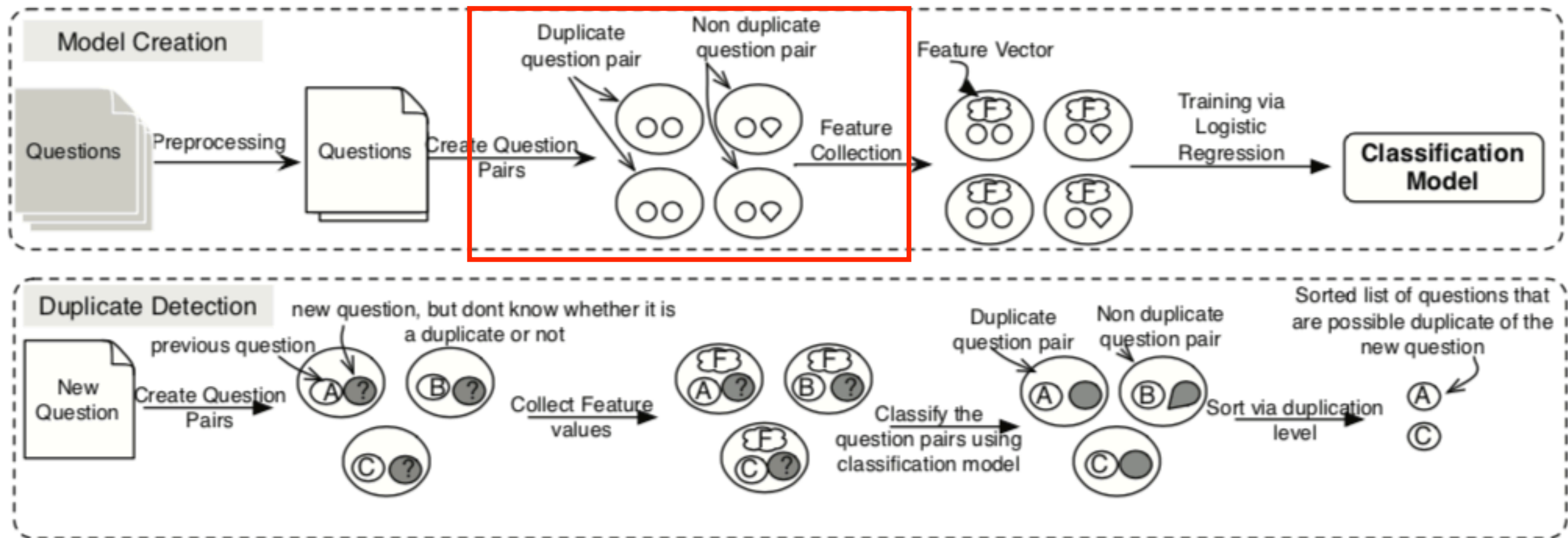


Figure 10: Classification model creation and duplicate detection via the model

- 質問を組み合わせることで、重複した場合のペアと重複していない場合のペアを作成し、特徴量を抽出したのちモデルを作成する

評価

- 他アルゴリズムとの比較
 - 既存手法であるDupePredictor, 関連手法BM25アルゴリズム
 - ・ BM25アルゴリズム：
 - ・ 文書におけるクエリの出現頻度を利用して関連度を順位づける
- 評価指標
 - Recall-Rate
 - ・ N_t : 重複だと判断した質問
 - ・ N_{dk} : 実際に重複であった質問

$$recall-rate_k = \frac{N_{dk}}{N_t}$$

結果

- Dupeが提案手法（80%を学習データ20%をテストデータ）

Table 3: Evaluation Results

Question Group	Technique	Recall (%)		
		Top-5	Top-10	Top-20
Java	BM25	28.33	31.14	35.12
	Dupe	38.25	44.55	53.02
	DupPredictor	30	35	41
	SO Search	19.03	26.91	-
C++	BM25	23.32	26.36	31.09
	Dupe	37.14	40.12	49.93
	DupPredictor	26	28	35
	SO Search	18.81	22.40	-
Python	BM25	26.10	30.18	36.09
	Dupe	37.20	45.41	53.22
	DupPredictor	28.15	32.28	38.74
	SO Search	17.21	25.41	-
Ruby	BM25	35.56	40.84	47.53
	Dupe	51.16	59.56	66.11
	DupPredictor	33	37	39
	SO Search	15.54	22.81	-
Html	BM25	23.94	27.11	31.22
	Dupe	37.14	39.45	50.23
	DupPredictor	25	28	34
	SO Search	16.83	26.04	-
Objective-c	BM25	26.81	31.61	37.11
	Dupe	40.61	47.88	56.35
	DupPredictor	26	28	31
	SO Search	17.73	25.52	-

特徴量ごとの精度

rubyの質問に関する調査

Table 4: Impact of different features on the recall-rate

No.	Different Source of Information	Top 5 (%)	Top 10 (%)	Top 20 (%)
1	Title (Cosine Similarity)	24.20	28.90	33.40
2	Title (Term Overlap) + Title (Cosine Similarity)	17.24	20.13	23.41
3	Body (Cosine Similarity)	24.50	27.50	32.80
4	Tag (Cosine Similarity)	18.00	24.00	26.00
5	Title + body + tag	34.50	40.10	47.20
6	title_body+body_title + title_tag	40.80	47.10	55.50
7	Step-5 + Step-6	45.80	54.10	63.50
8	Step-7 + code (<i>Dupe</i>)	51.20	59.40	66.10
9	step 8 +(Entity Overlap + Entity Type Overlap + WordNet Similarity)	51.40	59.63	66.35

まとめ・所感

- まとめ
 - StackOverflowにおける重複質問を検出する手法の提案を行った
- 所感
 - StackOverflowの質問が勉強になった
 - BM25, 特徴量で使っていた要素 (NER, WordNet Similarity) は勉強する価値がありそう