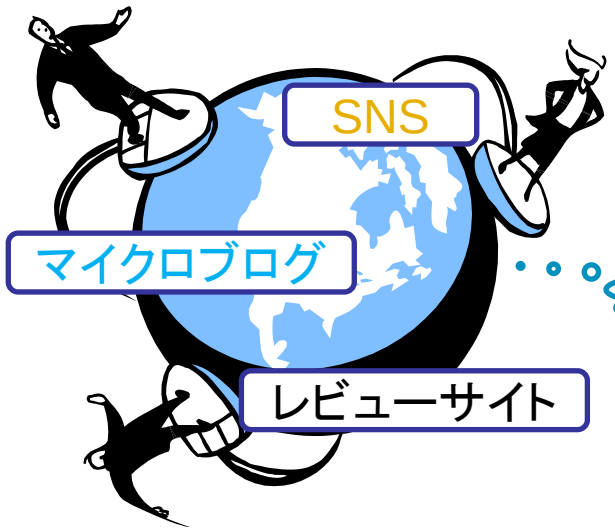
トピック推定に基づくソーシャルメディアからの耳より情報抽出手法の提案

甲南大学
服部　祐基　灘本明代

背景1



ソーシャルメディアを用いて情報を取得する
行為が増えている

コミュニティ

個人の経験談や体験談
ユーザ同士での情報交換

公式サイトや一般のWebページに掲載されていない
耳よりな情報が含まれている.

例えば

コミュニティ

諏訪湖花火大会

神奈川方面からなら諏訪インターの一つ前の諏訪南インターから下道が良いかと。やはり自動車よりも電車で行かれることをお勧めします.

コミュニティ

京都の紅葉

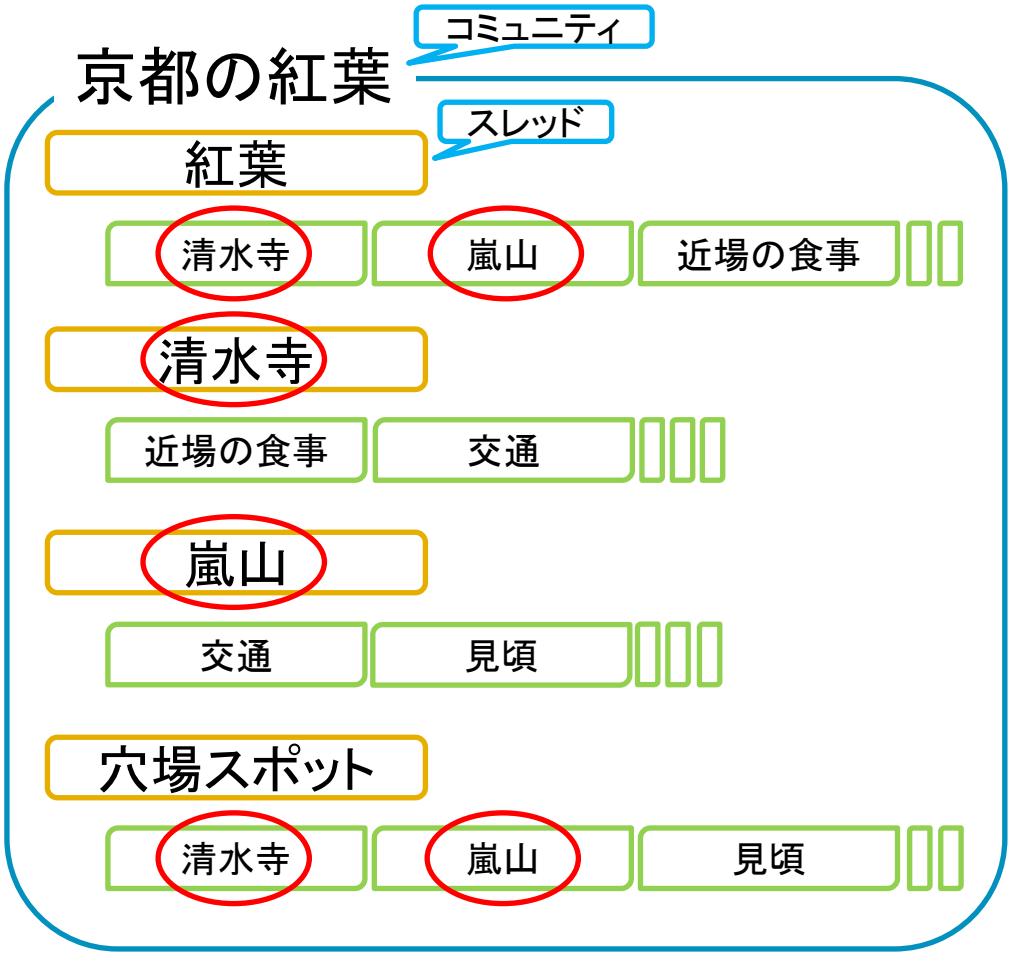
紅葉を見上げて見るなら洛西の「光明寺」、囲まれたいのなら、嵐山の「常寂光寺」、紅葉を見下ろすなら「東福寺」、紅葉の絨毯を見たいのであれば「永観堂」.

公式サイトや一般のWebページに掲載されていない
耳よりな情報

ソーシャルメディアから耳よりな情報を取得することが求められている.

背景2

SNSはコミュニティ内にスレッド（トピック）が作成されている.



問題点

様々な話題が混在.

ユーザが求める耳よりな情報
を取得することは困難.

目的

ソーシャルメディアから

トピック毎 に 耳より情報 を抽出

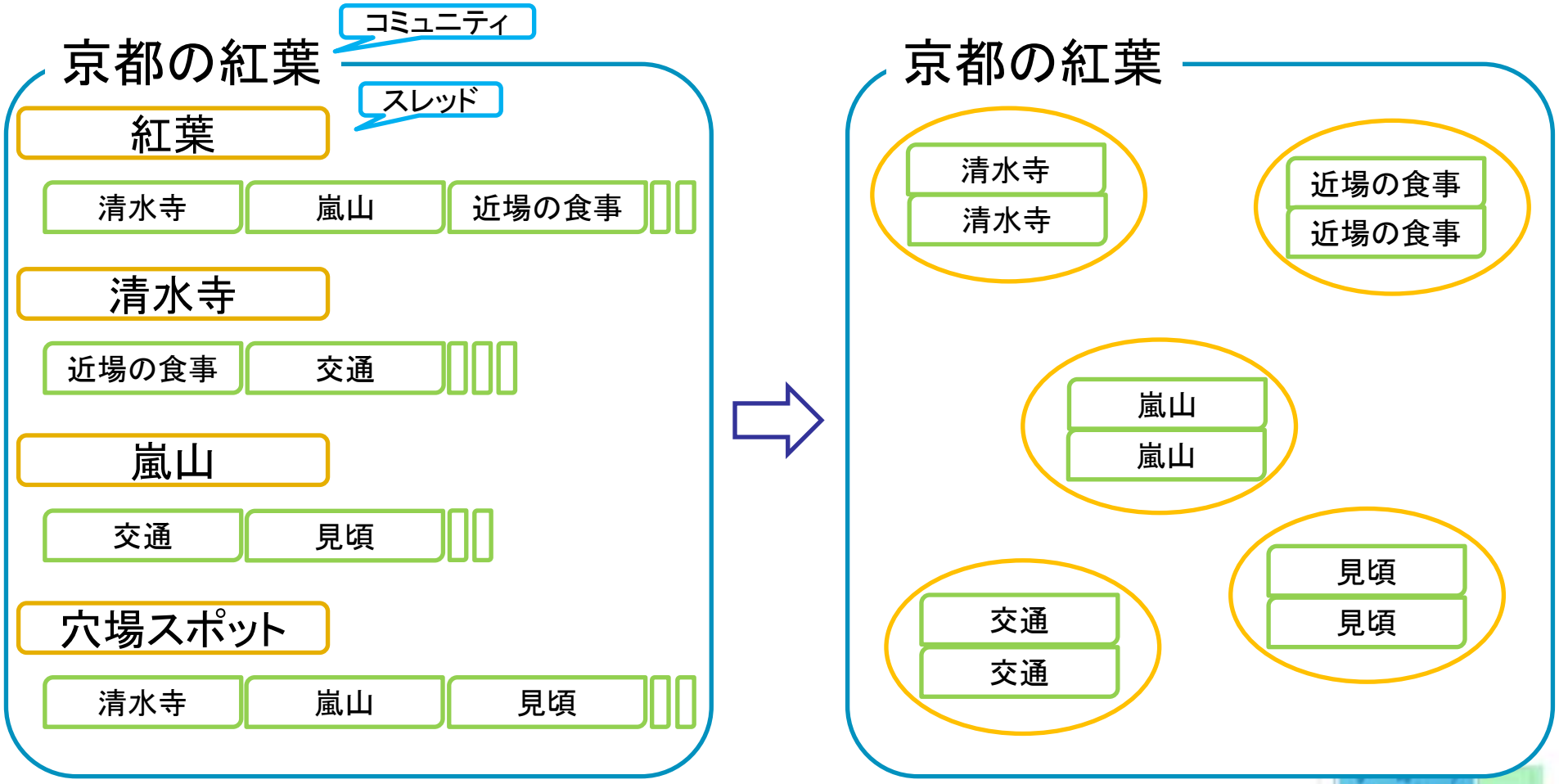
トピックの抽出と
クラスタリング

経験に基づいた
有益な情報

ユーザは効率的に有益な情報を取得することが可能となり，
新たな知識の獲得が期待できる

トピック抽出とクラスタリング

様々なトピックが混在しているため，コメントを1文書とし，そのコメント群からトピック推定を行うことにより，トピックの抽出，クラスタリングを行う．



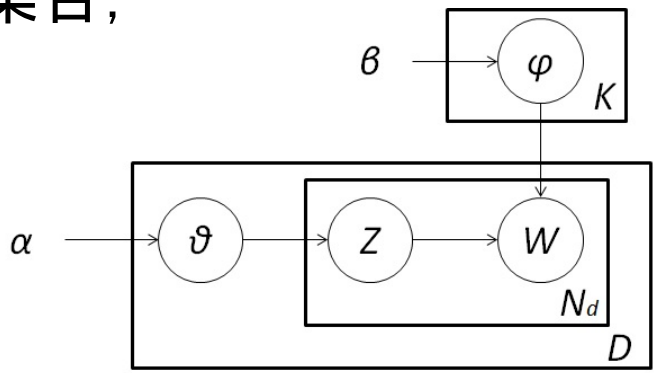
潜在的ディリクレ配分法（Latent Dirichlet Allocation, LDA）

潜在的ディリクレ配分法（Latent Dirichlet Allocation, LDA）

トピック推定手法

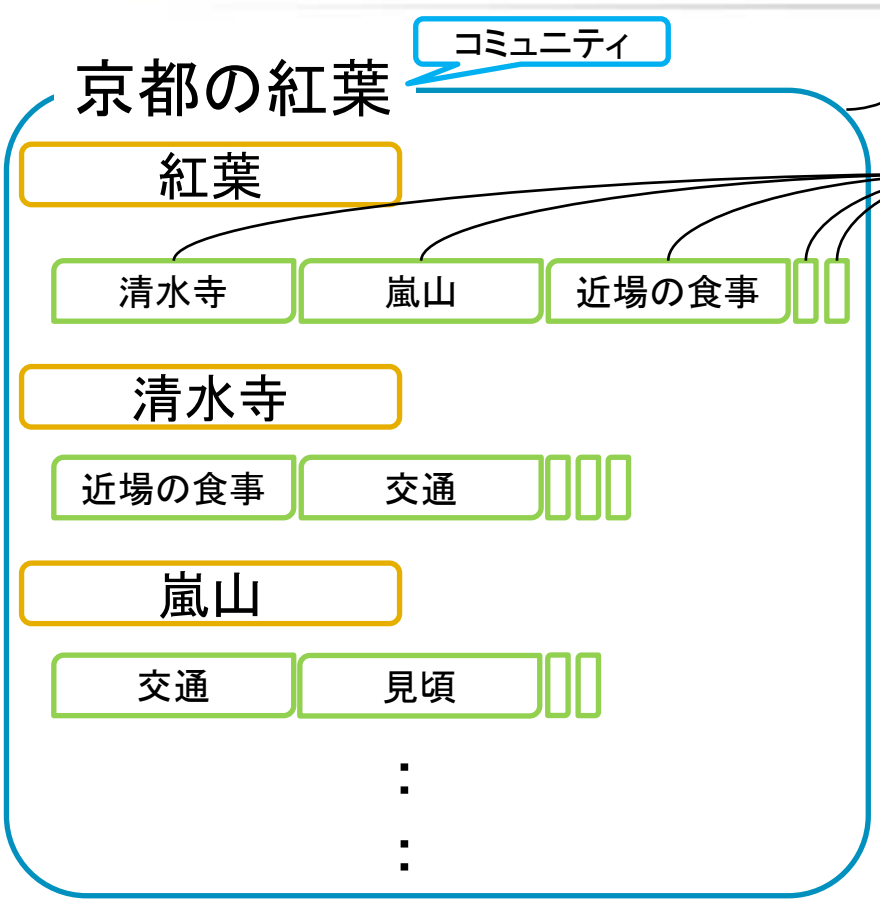
→ 文書集合からトピックを推定する際，データが疎な場合に有効

単語wの集まりにより表現された文書dの集合，
分類するトピック数K



各トピックzk (k = 1, ...,K)の単語w の確率分布P(w | zk)，
各文書d のトピックzk の確率分布P(zk | d) (d = 1, ...,D) を推定．

トピック抽出とクラスタリング



文書集合*D*

1コメントを1文書*d*

分類するトピック数*K*

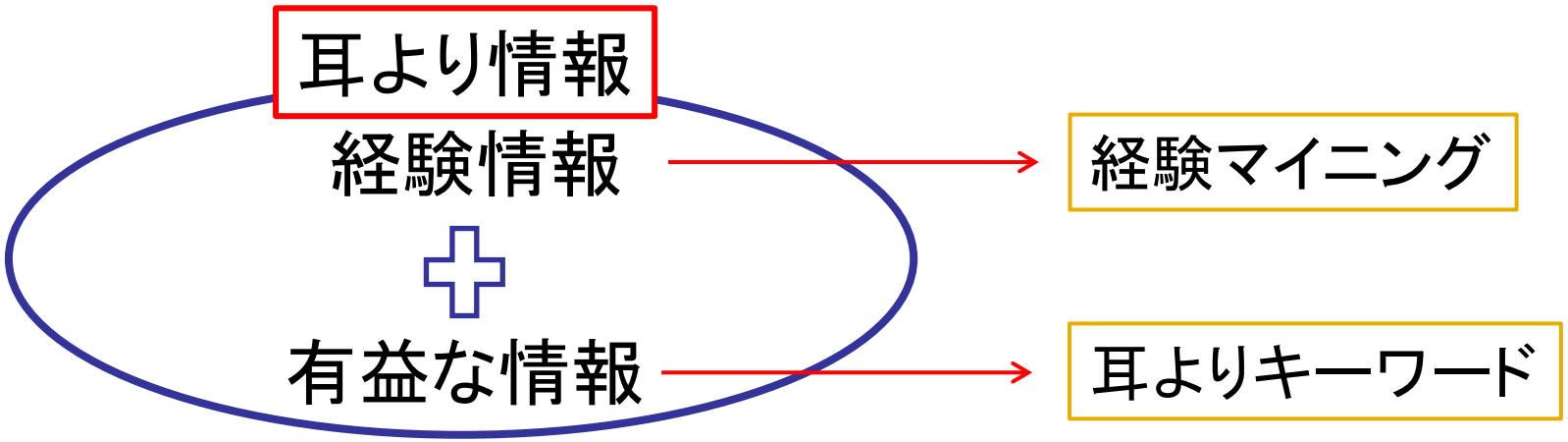
・トピック数*K*の決定（予備実験より）
5つのコミュニティに対してLDAによる推定
K=3, 7, 10, 15, 20, 25, …, 50.

適していた*K=10*を採用

文書*d*→トピック

各文書*d* のトピック*zk* の確率分布*P(zk | d) (d = 1, ...,D)*が0.3以上の値

耳より情報

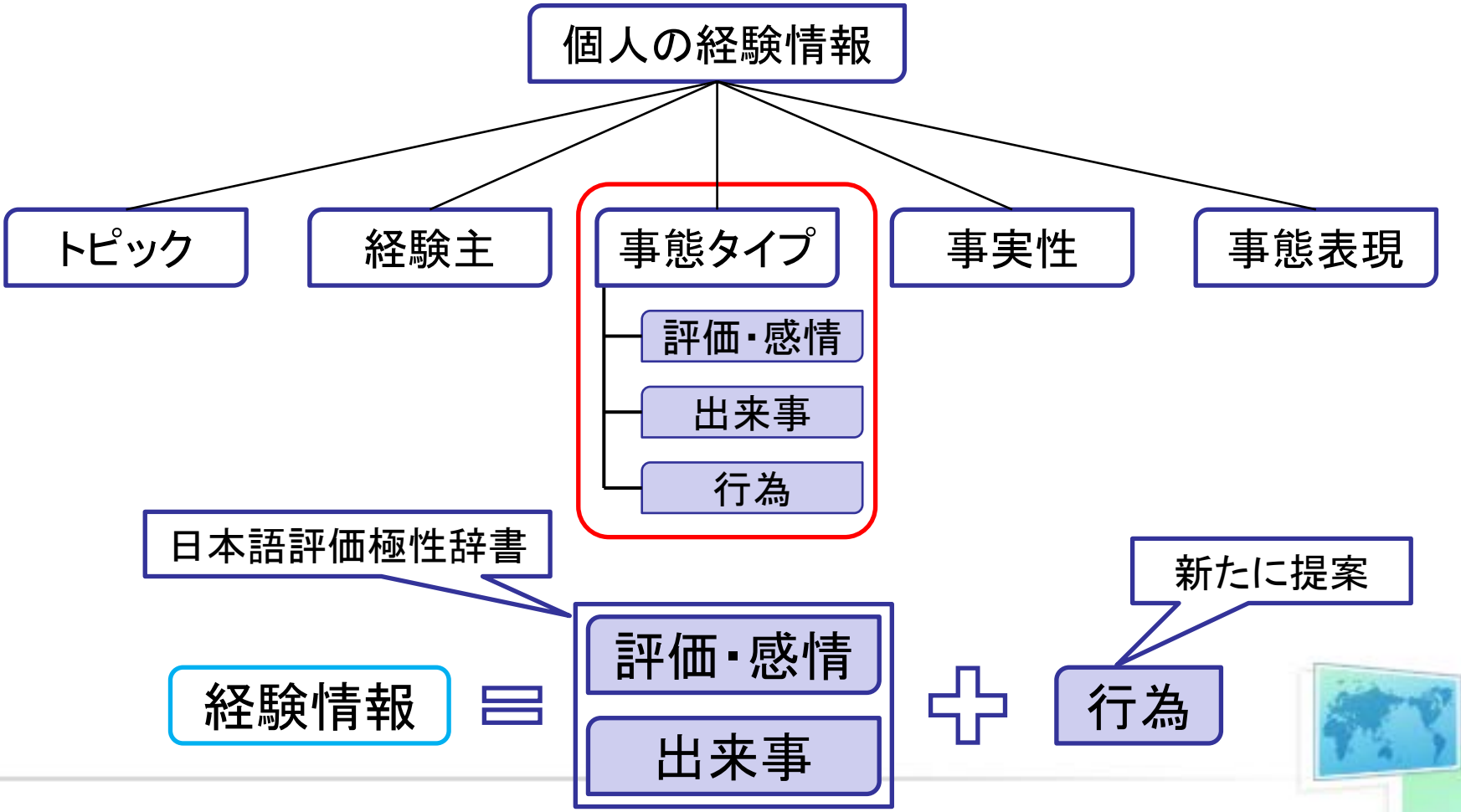


- 実際に経験したことに基づいたコメントは，より信頼できる．
- 多くの閲覧者が有益と感じた情報の中には共通のキーワードが含まれている．

経験情報の取得

経験マイニング Web上に大量にあるテキストから個人の経験情報を、構造化情報として抽出することを目的

乾健太郎，原　一夫：経験マイニング：Web テキストからの個人の経験の抽出と分類，言語処理学会第14 回年次大会論文集，pp. 1077–1080 (2008).



経験情報の取得

経験情報 ＝ 評価・感情 / 出来事 ＋ 行為

日本語評価極性辞書

動詞、形容詞、名詞に対してポジティブとネガティブの評価極性を付与している辞書

単語	タイプ
1番	状態
あいまい	評価・感情
あく抜け	出来事
いさかい	存在・性質
いたずら	行為
うってつけ	評価・感情
うだうだ	出来事

⇨ ・単語
・タイプ

経験情報の取得

経験情報 ＝ 評価・感情 / 出来事 ＋ 行為

経験マイニングでの行為 ⇨ 一般的な行為

我々の提案する行為 ⇨ ドメイン依存の行為

行為を表す単語の取得 ⇨ ドメインをイベント関連

イベント関連のコミュニティ30000コメント

⇨ 行為を表す名詞と動詞各出現頻度
上位50単語を取得

行為の例			
動詞		名詞	
行く	買う	参戦	利用
見る	並ぶ	参加	入場
できる	探す	乾杯	移動

経験情報を取得

取得した経験情報の例

・私も京都初一人旅へ行ってきました。2日間宿で自転車を借りて（ここは500円でした）とても効率よく回れました。盆地なので坂がきついところもありますのでなかなかいい運動になります。私は個人的に嵐山の大河原山荘が一番良かったです。庭園が美しかった・・。紅葉には早かったですが、満喫できましたよ。あとは左京区の曼殊院門跡もオススメです。もしそちらに行くなら詩仙堂も立ち寄るといいと思います。こちらもお庭が絶景です。

・高台寺、清水寺に行ってきました。綺麗な紅葉を見させていただきました。

・移動はやっぱレンタサイクルがオススメですよ．電車やバスだと時間を気にしながら行動しなくちゃならないけど、レンタサイクルなら自由。乗り捨てOKだったりするので、好きなところで借りて、好きなところで返せるっていうのも嬉しい

・秋に苔寺を見に行った時に[illegible]赤い紅葉があってとてもきれいでしたよ

青：行為
緑：評価・感情，出来事

耳よりキーワードの抽出

実際の耳より情報から耳より情報の要因となる
耳よりキーワードを抽出.

予備実験

被験者：5名

データセット：mixiのイベント関連5つのコミュニティ2000コメント

➢5人中4人が耳より情報であると感じたコメントを収集

⇨人手により耳より情報の要因となる耳よりキーワードを抽出

92種類

耳よりキーワードの例				
おすすめ	の方がよい	いかが	してみて	望ましい
すべき	是非どうぞ	できます	チャンス	イチオシ
狙い目	穴場	便利	お得	無難
ガラガラ	渋滞	混雑	絶景	最適

耳よりキーワードが含まれている情報を耳より情報とし抽出

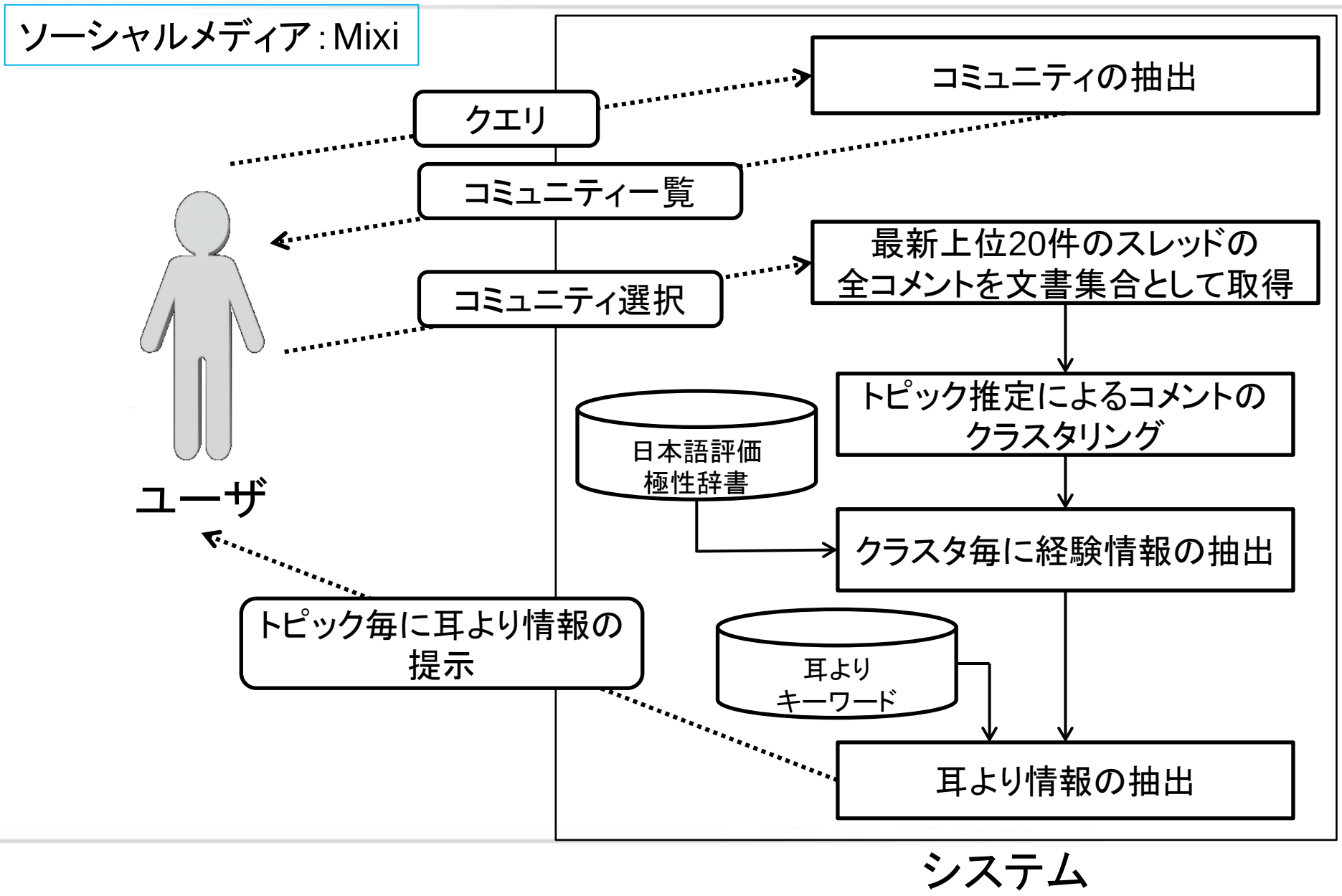
耳より情報の例

・私も京都初一人旅へ行ってきました。2日間宿で自転車を借りて（ここは500円でした）とても効率よく回れました。盆地なので坂がきついところもありますのでなかなかいい運動になります。私は個人的に嵐山の大河原山荘が一番良かったです。庭園が美しかった‥。紅葉には早かったですが、満喫できましたよ。あとは左京区の曼殊院門跡もオススメです。もしそちらに行くなら詩仙堂も立ち寄るといいと思います。こちらもお庭が絶景です。

・移動はやっぱレンタサイクルがオススメですよ．電車やバスだと時間を気にしながら行動しなくちゃならないけど、レンタサイクルなら自由。乗り捨てOKだったりするので、好きなところで借りて、好きなところで返せるっていうのも嬉しい

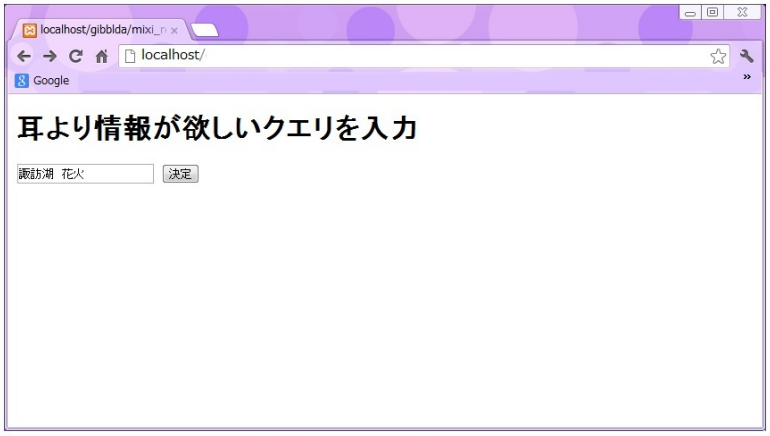
赤：耳よりキーワード
青：行為
緑：評価・感情，出来事

プロトタイプシステムの流れ



プロトタイプシステム

入力画面



コミュニティのリスト画面

コミュニティ名 →

トピック →

耳より情報のリスト →

出力画面

実験

- 目的:

  - ✓ 実験1: SNSのコメント群のクラスタリングにLDAが有効かどうか

  - ✓ 実験2:システムの有効性を測る
    - 提案手法とベースライン手法との比較
      - ↳ クラスタリングを行なっていない
    - 複数の被験者による精度の差異

実験1

◆ 目的：SNSのコメント群のクラスタリングにLDAが有効かどうか
一般Web文書に比べデータが疎であり，かつ文構造がきれいでないSNSのコメントに対するLDAの精度

◆ 実験条件：
- ✓ データセット： mixiのイベント関連の5つのコミュニティ

PL花火芸術　　なばなの里　　関西周辺海水浴遊び
京都紅葉巡り　潮干狩り

クラスタリングされたコメントとトピックを表すキーワードから，
コメントがトピックに関するものであるか

- ✓ 評価指標：適合率

実験1

結果

コミュニティ名	全コメント数	正しくクラスタリングされたコメント数	精度(%)
PL花火芸術	569	459	81
なばなの里	125	96	77
関西周辺海水浴	171	125	73
京都の紅葉巡り	162	124	77
潮干狩り	227	175	77
平均精度	--	--	77

- 一般Web文書に比べデータが疎であり，かつ文構造がきれいでないSNSのコメントに対して，良い結果が得られている
- 明らかに情報量が少ない場合では，正確に分類することができていない

実験2

◆ 目的：システムの有効性を測る
- 提案手法とベースライン手法との比較
- 複数の被験者による精度の差異

◆ 実験条件：
- ✓ 被験者：5名
- ✓ データセット：mixiのイベント関連の8コミュニティ

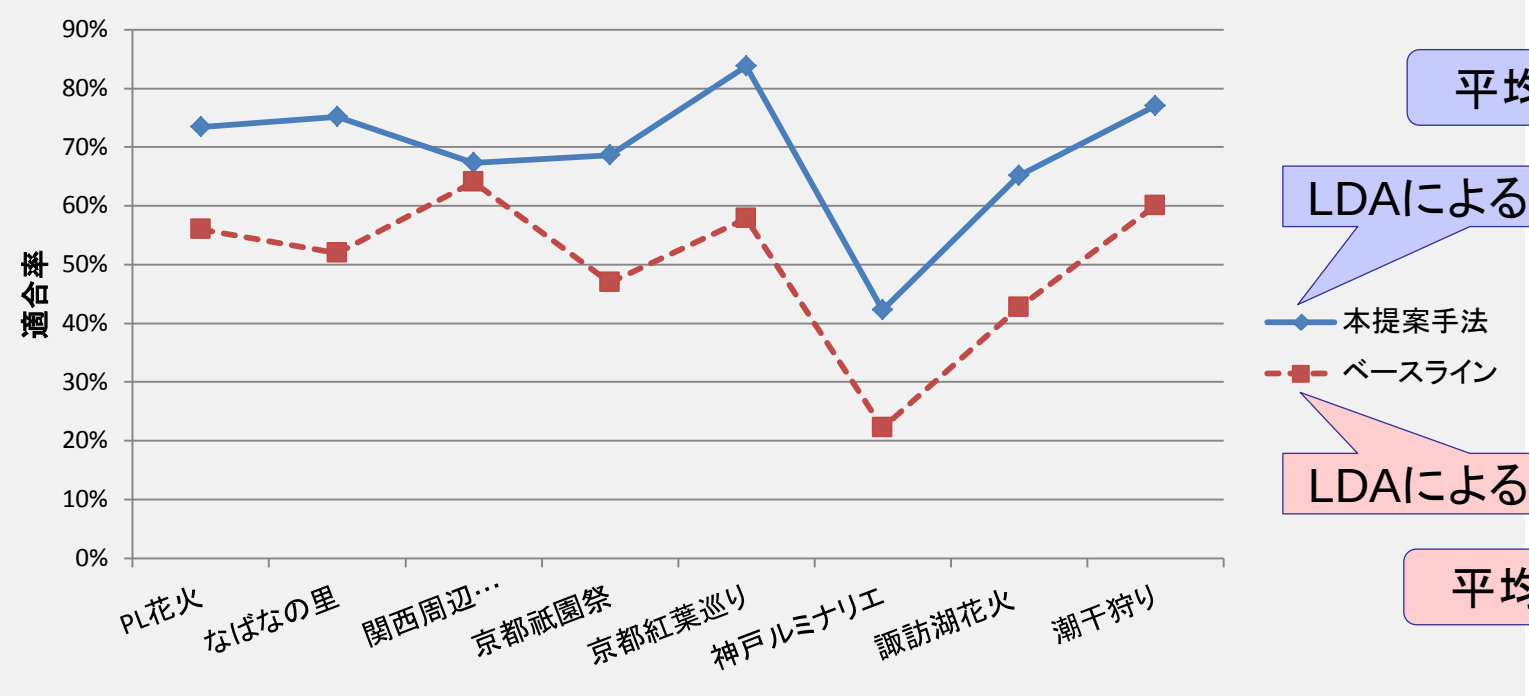
PL花火芸術	なばなの里	関西周辺海水浴遊び
京都紅葉巡り	潮干狩り	神戸ルミナリエ
京都祇園祭	諏訪湖花火大会	

システムが抽出した耳より情報のうち正しく抽出されたコメントの割合

- ✓ 評価指標：適合率

実験2：ベースライン手法との比較

結果

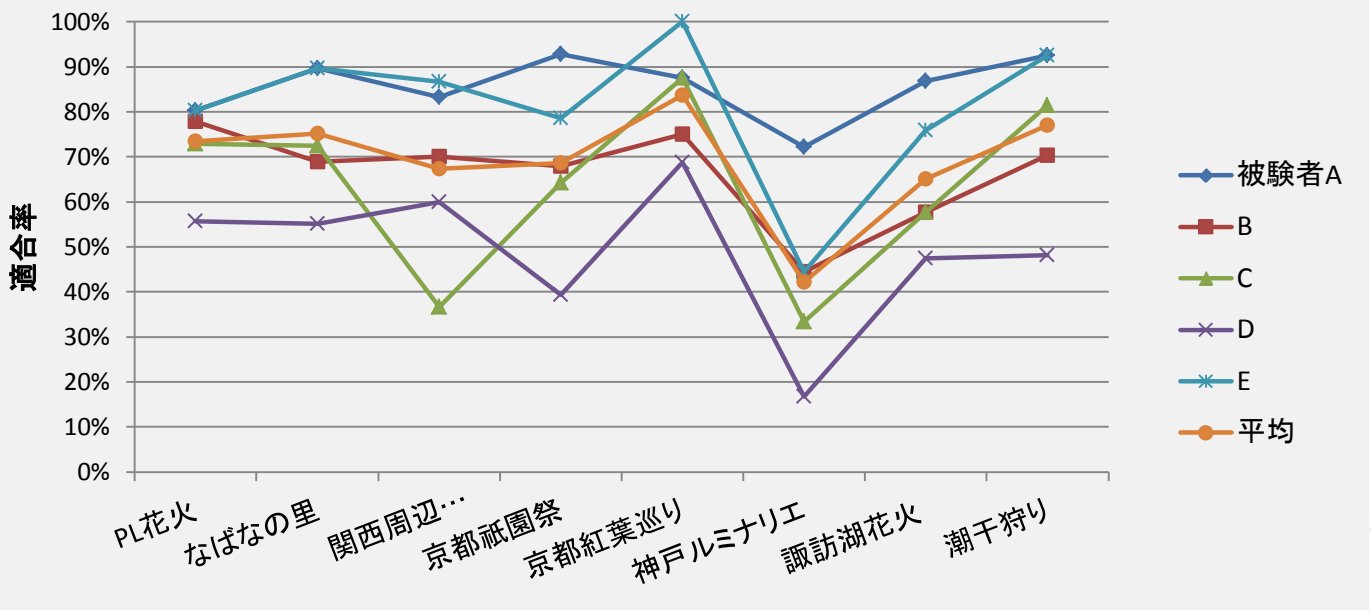


- 1つのコメントに対して耳より情報かどうかを判定している
  - → LDAによる分類を行わない場合，そのコメントが何についての話題か判断できないため，適合率が低い．
  - → LDAによる分類を行う事で，そのコメントにトピックが与えられ，話題を判断することができ，適合率が高くなった．

実験2：複数の被験者による精度の差異

結果

	PL花火芸術	なばなの里	関西周辺海水浴	京都祇園祭	京都紅葉巡り	神戸ルミナリエ	諏訪湖花火	潮干狩り	平均適合率
全コメント数	909	262	318	441	329	396	1313	415	-
取得されたコメント数	122	29	30	28	32	18	137	27	-
被験者A	80%	90%	83%	93%	88%	72%	87%	93%	86%
B	78%	69%	70%	68%	75%	44%	58%	70%	67%
C	73%	72%	37%	64%	88%	33%	58%	81%	63%
D	56%	55%	60%	39%	69%	17%	47%	48%	49%
E	80%	90%	87%	79%	100%	44%	76%	93%	81%
平均	73%	75%	67%	69%	84%	42%	65%	77%	69%



耳より情報の例

京都の紅葉

お勧めは高台寺。池に逆さに映る風景は息をのむ美しさでした！

東福寺は、凄い人ですが、紅葉も見頃を迎えると、通天橋から見下ろす景色は、順番に人ごみの中ならんでも価値ありの本当に凄く綺麗な景色です。

諏訪湖花火大会

諏訪市内に駐車してフィナーレまで見るなら諏訪インターまでは数時間の覚悟が必要です。岡谷駅周辺に駐車して電車で移動し帰りは岡谷インターから入る、または茅野駅周辺に駐車して諏訪南インターから入るという手段が最近流行っています。

有料自由席では必ず花火を間近で観ることは出来ます しかし、より観やすい場所にシートを広げて座りたいと思うのなら開場前から並ばなくてはなりません 速い人は何日も前から泊まり込んでいます 早朝から並んだとしても既に数百人の後ろです

まとめと今後の課題

ソーシャルメディアを用いて，トピック毎に耳より情報を抽出．

耳より情報
- ◆経験情報 → 経験マイニング
- ◆有益な情報 → 耳よりキーワード

LDAを用いてトピック毎に分類することで，ベースライン手法より１８%精度が向上

今後の課題

- 個々のユーザにあった耳より情報を提示
- 複数のソーシャルメディアから耳より情報の取得