NAIST-IS-MT0551037

修士論文

評価付けの重みを考慮した協調フィルタリング手法の提案と評価

川口 誠敬

2007年2月1日

奈良先端科学技術大学院大学
情報科学研究科 情報システム学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
修士 (工学) 授与の要件として提出した修士論文である。

川口 誠敬

審査委員：

砂原 秀樹 教授	（主指導教員）
山口 英 教授	（副指導教員）
藤川 和利 助教授	（副指導教員）

評価付けの重みを考慮した協調フィルタリング手法の提案と評価*

川口 誠敬

内容梗概

現在主流となっている全文検索システム上では、ユーザが目的に合ったキーワードを見出せないことや、検索結果数が膨大であることから、ユーザの嗜好に合った情報収集が困難になりつつある。そこで目的に特化した検索サービスが提供されている。このサービスはユーザが求めている対象の一般的に公開されている情報を得ることは可能だが、ユーザの嗜好に合うかという情報は提供されていない。なぜなら嗜好情報は、同じ表現や同じ文章、また同じ数値であっても、解釈する人によって捉え方が異なるため、取り扱いが困難な情報である。

このように、人の嗜好はバラつきがあり不安定な情報である。そのため、インターネット上に公開されている飲食店に対する情報を検索した場合、どの情報がユーザにとって有用であるか、つまり嗜好に合う情報かの判断が難しい。そこでユーザの嗜好を考慮した情報提供を行うものとして情報フィルタリングが注目されている。これらは不安定な情報である嗜好情報を、情報フィルタリングの技術である内容ベースフィルタリングや協調フィルタリングを利用することで嗜好に合った情報提供を実現している。ユーザの明示的な評価の数値入力は最もユーザの嗜好判断に適用しやすい情報であり、これをフィルタリングに利用し、より嗜好に合った検索結果が得られる。しかし、数値入力された評価値をそのままの値で利用しているため、ユーザがつけた本来の評価付けの重みが考慮されていない。そこで本研究では、ある対象に対して評価付けを行ったユーザの評価分布を利用

*奈良先端科学技術大学院大学 情報科学研究科 情報システム学専攻 修士論文, NAIST-IS-MT0551037, 2007年2月1日.

して基準点を決定する。基準点とは、ユーザにとって評価の良し悪しを決定する点である。基準点を用いることでユーザが評価付けした値に本来の重み付けが可能となる。ユーザ間の嗜好のズレを考慮し、嗜好に合った情報提供システムを提案する。提案手法に基づいて協調フィルタリングを活用した情報推薦システムを実装・評価した結果、既存技術より少ない誤り数でユーザの嗜好に合った情報を推薦することができた。

キーワード

情報推薦, 協調フィルタリング, 嗜好情報

Proposal and Evaluation of collaborative filtering method considering each weight of user's evaluation to restaurant*

Yoshihiro Kawaguchi

Abstract

Now, in widely the full-text search used, users become difficult to collect the information matched the preference of the user because they can't find the keyword matched the purpose or treat the flood of research results. So, Web service for satisfy user's special intention is appearing. This service provides the ordinary information that user request, but don't consider user's preference. It is difficult to treat the preference information which is generated from a huge amount of users. In this way, user's preference is unstable information for computer. That is why, when we search the restaurant information through the Internet, it is difficult to decide whether which information published is useful for me. So, the information filtering which considers user's preference is attracted. To treat the information of the preference, this method actualizes the providing information that we request by using content-base-filtering and collaborative filtering that it is technology of the information filtering. Inputting the number of the value explicitly is the information that is easy to apply to judge the user 's preference and use this to filtering, we'll get the search result that matched the preference more. But, related works don't consider the real weight of the value that weighted by user because it uses the untouched value input numerically. So, in this research,

*Master's Thesis, Department of Information Systems, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-MT0551037, February 1, 2007.

determine the standard value by using the user distribution estimated against something, The standard value is the basing point whether it is good or not for user. It actualize to treat the real weight of the value by using the standard value. I suggest the information proposedsystem that considers the preference gap of each user and can match the preference of the user. As a result of implement and evaluate recommendation system utilizing collaborative filtering based on proposed method, error counts of proposed method is less than that of relative works.

Keywords:

recommendation,collaborative filtering

目 次

1. はじめに **1**

1.1 背景 1

1.2 既存研究の問題点と本研究の目的 2

1.3 本論文の構成 3

2. 検索者の嗜好に合った情報検索 **4**

2.1 検索プロセス 4

2.2 必要情報の分類 5

2.2.1 安定したインデックス 5

2.2.2 不安定なインデックス 6

2.3 嗜好情報に関する関連研究 7

2.3.1 内容ベースフィルタリング 7

2.3.2 協調フィルタリング 9

2.4 要求事項 13

3. 提案手法 **15**

3.1 提案システムの概要 15

3.1.1 基準点の決定 17

3.1.2 評価値の重み付け 17

3.1.3 類似度計算 19

3.1.4 コンテンツの収集 19

4. 情報推薦システムの設計 **23**

4.1 システムの設計概要 23

4.2 システムの詳細設計 26

4.2.1 コンテンツの収集方法 26

4.2.2 DB へ格納 27

4.2.3 類似度計算 29

4.2.4 推薦結果表示 30

5. 情報推薦システムの実装 **31**
5.1 開発環境 31
5.2 利用画面例 31

6. 評価 **35**
6.1 評価環境 35
6.2 実験 1：基準点同定の性能評価 35
6.3 実験 2(実証実験)：類似ユーザの評価 38
6.4 実験 3(実証実験)：推薦店舗の評価 40
6.5 今後の課題 40

7. おわりに **42**

謝辞 **44**

参考文献 **46**

図 目 次

1 内容ベースフィルタリング概念図 8
2 協調フィルタリング概念図 10
3 本提案システムの全体図 16
4 「食べログ」ユーザトップページ 21
5 「食べログ」飲食店トップページ 22
6 提案システムのフローチャート 24
7 店舗評価登録前 32
8 店舗評価登録後 33
9 オススメ検索結果の表示画面 34
10 嗜好に合ったユーザの割合 39
11 推薦店舗に対する被験者の嗜好の一致性 41

表 目 次

1 メモリベース法とモデルベース法の比較図 12
2 飲食店の基本情報テーブル 28
3 ユーザテーブル 28
4 ユーザの飲食店に対する評価テーブル 28
5 基準点テーブル 29
6 開発環境の詳細 31
7 評価環境 35
8 各実験の被験者数 35
9 検証結果例 36
10 異なる基準点毎の誤り数 37
11 異なる基準点毎の情報推薦数 38

1. はじめに

本章では、本研究を取り巻く背景とそこに存在する問題点を挙げ、本研究の目的を示す。

1.1 背景

近年、Web 上の情報源の急速な膨大化、ユーザ数の増加に伴い、その利用目的も多様化している。そのため利用目的に合わせた情報提供サイトが発展してきた。またプロバイダの提供する簡易 HP(Home Page) 作成サービスや、SNS(Social Networking Site)、ブログ (Weblog) を使って、誰もが容易に情報発信可能となった。そのため、利用目的毎にさらに情報提供サイトが作られてきている。代表的な検索技術であるキーワード検索のみでは、条件の絞込みが甘かったり、キーワードが思いつかない、また同一のキーワードを入力した場合、検索結果のランキングが利用しているユーザ毎に変化せず、ユーザの嗜好に合致した情報を得ることが困難になっている。そこでユーザの嗜好などの情報を登録させ、ユーザの嗜好に合致した情報を提供する仕組みとして、情報フィルタリングが注目されている。例えば、代表的な情報フィルタリングを用いた推薦システムとして Amazon.com [1] のような書籍推薦システム、TSUTAYA online [2] のような映画や DVD の推薦システムが挙げられる。

一方、収集する情報源にも変化が現れている。今までは企業や利用目的に応じた大手サイト（以後、既存サイト）が情報提供者となり、事業主から直接得る最新情報を基に Web サイトを構築することなど、信頼できる正確な情報を提供していた。しかし、既存サイトでは、掲載されている情報が提供者側の不利益にならないような情報しか掲載されていない。そのため、掲載情報からは、ユーザが本当に知りたい評価・評判情報を得ることが困難であり、ユーザの嗜好に合致する情報か、といった各ユーザにとっての情報の良し悪しを判断するには不十分な情報である。そこで、消費者が発信する情報が注目されている。インターネット

[1] Amazon:http://www.amazon.co.jp/
[2] TSUTAYA online:http://www.tsutaya.co.jp/index.zhtml

の普及とSNSやブログといったコストもほとんどかからず、技術的な知識を必要とせずに誰でも容易に情報発信できる仕組みができたことにより、既存サイトでは掲載できないような評価・評判情報が発信されるようになってきた。また、既存サイトとは別に口コミ情報を中心として収集するような口コミサイト (価格.com [3]、食べログ.com [4]) も登場し、多くの一般人が記述した評価・評判情報を閲覧することが可能となった。情報発信者が一般の個人であることから、情報発信する内容も様々である。SNSやブログには日常生活を日記として記述している人は多い。その中でも誰しもの私生活に欠かせない飲食に関する情報を日記という形で残している人は多い。また飲食店へ行った時に、出される料理の写真を撮り、口コミ情報を収集しているブログサイト (食べログ.com) に書き込むことで、情報提供料として得られるポイントを集めているユーザもいる。このような一般の個人から発信されている飲食店に関する評価・評判情報を閲覧して、「ある飲食店に対して自分と同じ評価をしている人は嗜好が似ているため、その人の推薦する飲食店ならば自分も満足するであろう」という考えを無意識に利用することで、自分の嗜好に合致した情報を収集するユーザが増えつつある。

以上のように、嗜好に合致した情報を収集するには、口コミ（評価・評判）情報を含む情報源を閲覧し、嗜好の類似した情報発信者を発見することが重要となる。

1.2 既存研究の問題点と本研究の目的

前節で説明したように嗜好が似ているユーザ（以後、類似ユーザ）は、検索しているユーザの嗜好に合致する情報を公開していると考えられる。つまり類似ユーザの情報は嗜好が合致する可能性が高い。そこで、ユーザの嗜好に合致する情報を取得するためには、ある情報に対して「自分と同等の評価を下しているユーザのオススメ情報は自分にとってもオススメである」という協調フィルタリングの情報推薦モデルを利用して嗜好の似ている類似ユーザの発見を行う。その類似ユーザから推薦情報を取得する必要がある。既存研究では、嗜好の近さを測る目安としてユーザ間の類似度を計算することが一般的に用いられている。まず、明示的

[3]価格.com:http://kakaku.com/
[4]食べログ.com:http://r.tabelog.com/

にユーザにある情報に対して好きか嫌いか、興味があるか興味がないかといった嗜好情報を数段階で評価付けさせ、この作業を繰り返し行う。ある程度評価情報が蓄積したら、情報に対しての評価付けの傾向が似ているかどうかで類似度計算を行い、類似度が高い類似ユーザから推薦情報を取得する [1]。しかし、既存研究では、嗜好が似ていることに関して評価値が近いことしか考慮されておらず、評価値の重み付けまで考慮していない。そのため、ユーザによって評価付けの重みが異なるにも関わらず、評価値が等しいため重みも等しいものと解釈し、ユーザ間に嗜好のズレが生じてしまう。

そこで本研究では、各ユーザによって異なる評価値本来の重み付けを考慮した類似度計算を提案する。これによって嗜好の類似性を計算するときに、ユーザによって異なる基準点でも嗜好の近さを測ることができる。類似ユーザからユーザが高い評価を下すと予想される情報、満足させる情報を取得できることを本研究の目的とする。また本研究では、ブログや口コミサイトでの個人の情報発信を背景に、多くの人々の関心でもあるグルメ情報を推薦対象にする。

1.3 本論文の構成

2 では、嗜好に合致する情報取得に必要な手順、情報を述べ、既存技術の紹介を行う。また既存技術を考察することにより本研究の要求事項を明確にする。

3 では、要求事項を踏まえ、評価付けの重みを考慮した情報推薦システムを提案する。

4 では、提案したシステムの具体的な設計について述べる。

5 では、設計に基づいて実装を行った開発環境と、その詳細について述べる。

6 では、本研究のシステムの評価を考察する。

7 では、本研究の結果により明らかになった結論と今後の課題について述べる。

2. 検索者の嗜好に合った情報検索

本章では、まず、検索者の嗜好に合う検索結果を得るために一般的に取られている検索プロセスを述べる。次に、その検索に必要とされる情報とその性質を述べ、検索者の嗜好を考慮した情報提供に関する研究について述べる。最後に検索者の嗜好に合った情報検索に必要とされる情報の性質を踏まえて、本研究の目的を果たす情報推薦システムの機能要件を述べる。

2.1 検索プロセス

グルメ情報を収集する検索者は、一般的にグルメ検索サイトを利用している。グルメ検索サイトとは、「ぐるなび[5]」、「グルメぴあ[6]」に代表される飲食店の情報を集めたWebサイトであり、飲食店の情報を事業主から広告として募り、利用者は無料でグルメ情報を検索・閲覧できるサイトである。検索者の嗜好に合った情報提供を実現するために、位置に基づく検索や、目的、予算、料理のジャンル、内装の写真など様々な検索機能を提供することで検索者を満足させる検索結果を提示している。一方、個人が発信するグルメ情報も参考にされつつある。ブログやSNSの普及により、個人が容易に情報発信できるようになった。このことから、検索者は、飲食店の事業主が発信するような位置、平均予算、電話番号などといった飲食店の基本情報とは異なった、個人が発信する飲食店に対する口コミ情報（評価）を閲覧できるようになった。インターネット上に公開されている飲食店に対しての複数の口コミ情報を閲覧することで客観的に飲食店を評価できるようになった。

さらにある飲食店に対して検索者と同様の評価を下している個人を、嗜好の似ている個人とし、検索者の未開の飲食店に対してその個人が高い評価を下している場合、検索者も同様に高い評価を下すと考えることができる。これにより、検索者は嗜好に合った未開拓の飲食店情報を取得することができる。上述したように、検索者は、要求している情報によって異なるWebサイトを利用することで嗜

[5] ぐるなび:http://www.gnavi.co.jp/
[6] グルメぴあ:http://g.pia.co.jp/

好に合った飲食店情報を収集している。次節では、嗜好に合った飲食店情報を収集するために必要な情報とその性質を述べる。

2.2 必要情報の分類

前節で述べたように、様々な嗜好を持った検索者を満足させる検索結果を得るには、検索対象である情報源が以下に挙げる情報を含んでいることが望ましい。

- 店舗名、住所、電話番号、予算、料理ジャンル、メニュー、営業時間、休日、設備、飲食店公式 HP
- サービス、雰囲気、味の評価

情報を性質で分類すると、店舗名、住所、電話番号、予算、料理ジャンル、メニュー、営業時間、休日、予算、設備、飲食店公式 HP といった飲食店固有の情報は、どの検索者にとっても普遍な情報のため安定したインデックスと考えられる。またサービス、雰囲気、味の評価は事業主が広告として宣伝している情報以外にも個人が発信している情報も含まれ、価値観により内容が異なるため不安定なインデックスだと考えられる。

2.2.1 安定したインデックス

安定したインデックスは、どの検索者にとっても普遍な情報であり、目的の飲食店を決定する上では最低限必要な情報である。この情報は記述形式が共通であるため、構造化されている既存サイトや口コミサイトから容易に抽出できる情報である。その中でも特に既存サイトは、事業主が広告として情報を登録しているため、情報量の豊富さ、情報の新しさ、構造化された Web サイト、整理された情報表示などの理由から信頼のおける情報源である。このような安定したインデックスに関しては、口コミサイトや個人ブログよりも既存サイトから収集した方が多くの必要情報を得ることができる。

2.2.2 不安定なインデックス

不安定なインデックスは、受けたサービスの感想、店の雰囲気、味の評価といった、飲食店に関する評価情報である。既存サイトでは、個人によってバラつきのある情報は飲食店の評価をあいまいにさせ、事業主に不利益を与える恐れがあるためほとんど掲載されておらず、当たり障りのない均一な評価が掲載されていることが多い。この情報は、実際に経験した当事者が発信した情報の方が真実味があり、口コミサイトや個人ブログに投稿されている情報を収集することが望ましい。記述形式は情報発信者が個人であることから情報にバラつきがある。媒体が個人ブログの場合、利用しているサービスによってサイト構造が異なり、発信している情報に評価情報が含まれているかの判断、記述している箇所を特定し抽出することが困難である。また、口コミサイトや個人ブログにある飲食店に対して、「美味しかった」という記述がある場合を考えてみる。個人ブログに「美味しかった」と記述してある場合、その飲食店に対する評価は 10 段階評価に置き換えると 7 点かもしれない。また、別の個人がそのブログを閲覧した場合は「美味しかった」を 9 点と解釈するかもしれない。このように個人によって基準点が異なるため、文章のみでは個人の評価付けの重みを判断することは困難である。そのため検索者が口コミサイトや個人ブログから嗜好に合致する情報を収集することは困難である。そこで、不安定なインデックスである人の嗜好を取り扱う研究が行われている。

本研究では、この不安定なインデックスである飲食店の味に対する評価情報に着目する。情報発信者によって評価にバラつきのある評価情報に、基準点を設けることで検索者と他の情報発信者との嗜好の比較をより正確に行うことを可能とし、嗜好の類似するユーザを発見する。そのユーザからある検索者にとって未開である高い評価が予測される飲食店情報をオススメ情報として提示することを目的とする。

2.3 嗜好情報に関する関連研究

膨大な情報源の中から検索者の嗜好に合った情報を見つけ出し、情報を提供する情報推薦技術である情報フィルタリングが注目されている。情報フィルタリングは膨大な情報源から情報を検索者に提示する前に情報と検索者との関連性を計算する。その関連性を基に検索者が興味を持つような情報だけを抽出し検索者に提示することができる。この情報フィルタリングを実現するために主に用いられている手法としては内容ベースフィルタリングと協調フィルタリングの 2 つがある。内容ベースフィルタリングでは検索者の嗜好情報をユーザプロファイルとしてコンテンツの属性情報や情報に対するキーワードなどとのマッチングを行い、推薦するコンテンツを提示する。協調フィルタリングでは同様の嗜好を持ったユーザを発見し、その類似ユーザが高い評価を下した情報やコンテンツを推薦する。

2.3.1 内容ベースフィルタリング

内容ベースフィルタリング (content-based-filtering) とは、検索者の嗜好情報をプロファイルとして表現し、コンテンツとのマッチングを行い、検索者に合ったコンテンツを推薦する手法である。この手法の概念図を図 1 に示す。まず、ユーザ A の飲食店情報が掲載されている Web ページを閲覧する。自分の嗜好に合致する Web ページを閲覧する動作を繰り返し行うことで、Web ページ内に含まれている単語のベクトルからユーザの嗜好を表すユーザプロファイルを作成する。図 1 の例では飲食店に関する雰囲気、味、サービスといった評価値によって、嗜好のベクトルを形成している。このユーザプロファイルベクトルと一致したベクトルを持つ飲食店は飲食店 3 と飲食店 5 になる。そこで飲食店 3 と飲食店 5 をユーザ A に推薦すると、ユーザ A の嗜好に合致する可能性が高い。内容ベースフィルタリングでは、ユーザプロファイルと対象のコンテンツをいかに的確に定量化できるかが性能に影響を与える。検索者の嗜好を把握する研究は、サービスを提供する上で直接ビジネスチャンスにもつながることから多くの企業の研究機関で行われている。ユーザの Web 閲覧時の行動から、ユーザの嗜好情報を取得する方法の例をいくつか挙げる。閲覧したページのすべてに検索者が興味を持ったと仮

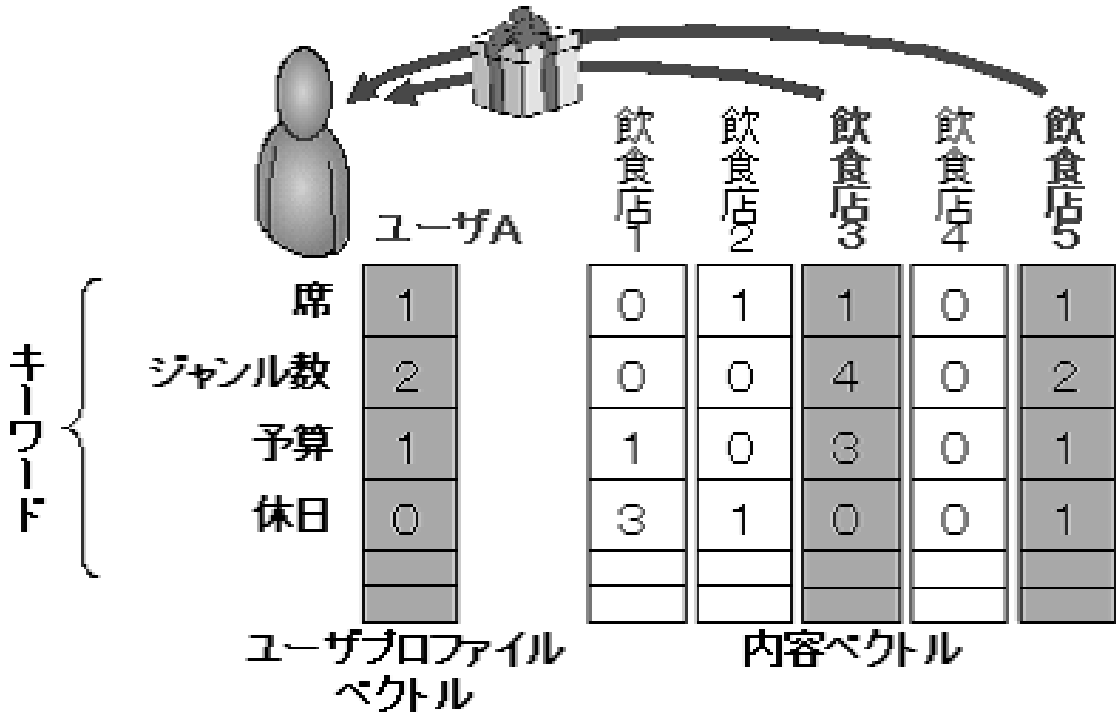

図 1 内容ベースフィルタリング概念図

定して、Web ページのアクセス履歴（Web サーバまたはプロキシサーバなどから得る「どのページを閲覧したか」という履歴）を用いる方法と、何らかの方法で検索者が閲覧した情報に興味があったか否かを判定する方法の 2 種類に分けられる。後者において、閲覧した情報に対する興味の有無を推定する研究をいくつか例に挙げる。例えば、検索者が閲覧に費やした時間と検索者にとっての記事の有用さの度合いとには相関関係があることを示し、閲覧時間からユーザプロファイルを作成する研究 [2] がある。また、閲覧中の検索者のマウス操作（ページに対する拡大表示ボタンを押したか否か、スクロールをしたか否か、拡大表示とスクロール両方を利用したか否か）からこれらの操作があったページに対して重み付けをしてユーザプロファイルを作成したり [3]、閲覧中の視線を利用する研究には、視線は左から右への跳躍運動が繰り返され、その行を読み終わると一気に次の行先頭へ移動する、といった行動モデルを基本に跳躍の幅を閾値として文章への注目度を検出し、注目度からユーザプロファイルの作成を行うもの [4] がある。またコンテンツの定量化に関しては、各ページが対象に対してどのような評価ポイントがあるのかを抽出し、それらのポイントについて各ページでどの程度言及されいてるかの尺度化を行った研究がある [5]。内容ベースフィルタリングの利点としては、cold-start 状況でも推薦対象の内容さえ分かれば、適切な推薦が可能なことが挙げられる。cold-start 状況 (cold-start condition) とは、システムを使い始めたばかりのユーザへ推薦したり、新しくシステムに登録されたものを推薦対象にする状況のことである。他に、推薦候補として考慮されるものの範囲が広く、少数派の嗜好の検索者でも比較的よい推薦が受けられる。内容ベースフィルタリングでは、飲食店がどのような評価を得ているかまでは考慮されず、飲食店に対する嗜好情報は必要情報である安定したインデックスのうちの席の数、ジャンル数、予算、休日といった情報にしか有効ではない。

2.3.2 協調フィルタリング

協調フィルタリング (collaborative filtering) とは、複数の類似したユーザのプロファイルや過去の行動履歴から新たな推薦すべきコンテンツを導出する手法である。この手法の概念図を図 2 に示す。過去に行ったことのある飲食店に対する

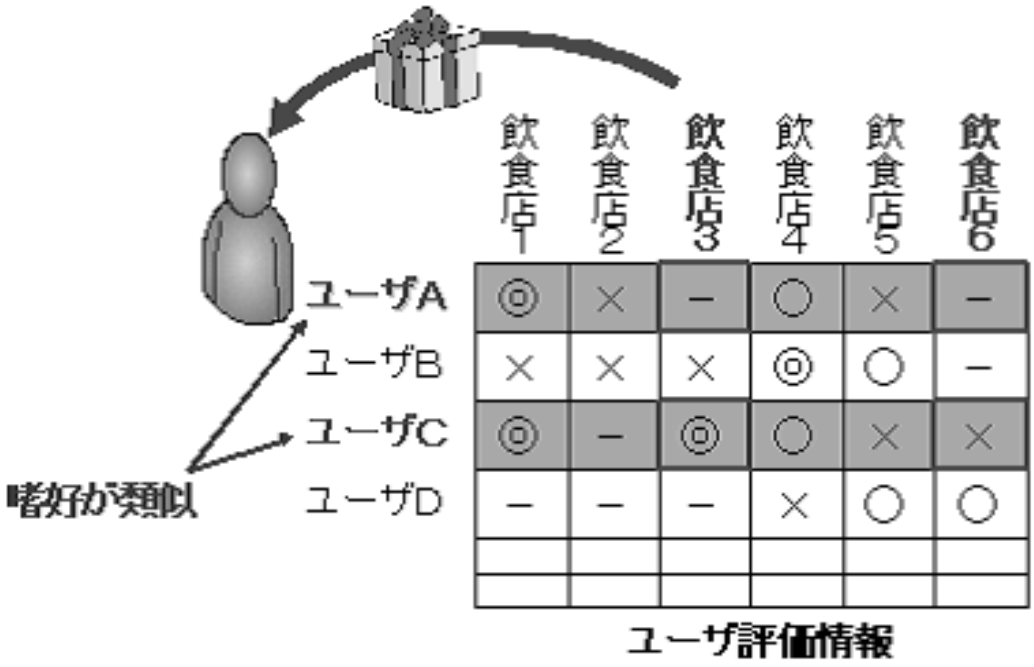

図 2 協調フィルタリング概念図

評価の傾向がユーザ A とユーザ C で類似している。このとき、ユーザ C が高く評価した飲食店 3 は、ユーザ A の嗜好に合致する可能性が高い。ユーザ A に飲食店 3 を推薦することによって、ユーザ A は嗜好に合致した飲食店を発見することができる。このように、協調フィルタリングには自分では知らないような飲食店に対しての推薦を類似ユーザと嗜好を比較することで、意外性のある推薦を受けられる利点がある。内容ベースフィルタリングでは、推薦対象の情報を利用するため、どうしても検索者が予想できる範囲の推薦になりがちである。一方、協調フィルタリングでは、他人の知識をうまく利用している。また、全体的に情報が少ない状況でも比較的適切な推薦ができることや、推薦対象の内容を確認しているわけではないため、情報収集の手間がないことが挙げられる。説明してきたとおり、協調フィルタリングでは検索者の嗜好情報である飲食店に対する評価値に基づいてシステムを利用する検索者の嗜好を予測する。検索者から得た嗜好情報の中から、規則性を見つけ出し、その規則性に基づいて予測するこのような問題は、機械学習や統計的予測によって解く。しかし全てに適応できるような予測手法は困難なため、検索者数や対象数などの利用者の性質や、推薦の利用目的に応じた手法が必要になり、様々な手法が開発されている。予測手法はメモリベース法 (memory-based method) とモデルベース法 (model-based method) に分けられる [7]。

メモリベース法では、推薦システムが利用される以前には何もせず、推薦対象を評価しているユーザとユーザの嗜好情報である評価値がユーザ DB(Data Base) として保持されている。そして、推薦をするときには、検索者の嗜好情報とユーザ DB 中の嗜好情報とを用いて予測をする。一方、モデルベース法では、推薦システムが利用される以前に、あらかじめモデルを構築する必要がある。モデルとは、「A さんが好きなものは、B さんも好きなことが多い」といったユーザと推薦対象の嗜好についての規則性を表したものである。推薦をするときには、ユーザ DB を用いずにこのモデルとユーザの嗜好情報とに基づいて予測する。

これらの手法の長所と短所を表 1 にまとめる。推薦時間は、メモリベース法の方が一般的に遅い。これはユーザ DB には多くのユーザや推薦対象が格納されており、多くの項目を推薦の度に走査するのは時間がかかるためである。モデルベー

表 1 メモリベース法とモデルベース法の比較

	メモリベース法	モデルベース法
推薦時間	×：遅い	○：早い
適応性	○：あり	×：なし

ス法では、モデルの事前構築に時間を要するだけであり、このことは推薦の速さに影響しない。また、モデルの規模は、ユーザ DB のそれと比べて小さいので、検索者に早く推薦することができる。表 1 の適応性は、ユーザ数や推薦対象数が変化しても適切な推薦ができるかということである。モデルベース法は、このような変化が生じるとモデルを再構築する必要が生じ、時間がかかるため適応性には長けていない。一方、メモリベース法は、モデル構築を行わないためこのような問題は生じない。メモリベース法の代表的なシステムとして、GroupLens [1] が挙げられる。GroupLens は、NetNews の記事の中から、検索者が関心をもつ記事を推薦するシステムとして開発された。だが、現在では NetNews があまり利用されなくなったことから、同じ手法を用いた映画の推薦システム MovieLens となっている。GroupLens の方法は次の二段階で実現する。

1. 類似度の計算：ユーザ DB 中の各ユーザと検索者との嗜好の類似度を求める。類似度とは、嗜好の傾向がどれくらい似ているかを数値化したものである。

2. 嗜好の予測：検索者が知らないが、ユーザ DB 中のユーザは知っている対象について、検索者がどれくらいその対象に関心があるかを予測する。

この過程で使われる類似度の計算式を以下に示す。

$$r_{AB} = \frac{\sum_i (A_i - \overline{A})(B_i - \overline{B})}{\sqrt{\sum_i (A_i - \overline{A})^2}\sqrt{\sum_i (B_i - \overline{B})^2}} \tag{1}$$

r_{AB}：ユーザ A とユーザ B の類似度
A_i：ユーザ A の対象 i に対する評価値
$\overline{A}$：ユーザ A の対象に対する評価の平均

B_i：ユーザ B の対象 i に対する評価値

このように、GroupLens は相関係数を類似度として用い、嗜好が近い類似ユーザを発見している。この場合の基準点はユーザ毎の評価値の平均を用いている。しかし、この平均値はユーザが評価した対象全ての評価値に対する平均値ではなく、ユーザ間の比較に用いられている評価値（共通の対象に対する評価）の平均を用いている。そのため、比較数が少ない場合は正しい評価値の重みが考慮されずに嗜好のズレが生じることが考えられる。

また ringo [6] はユーザが音楽 CD に対して評価付けを行い、蓄積した嗜好情報と利用者ユーザとの嗜好情報の比較から推薦を行う。以下のような式でユーザ間の類似度を算出している。

$$r_{AB} = \frac{\sum_i (A_i - 4)(B_i - 4)}{\sqrt{\sum_i (A_i - 4)^2}\sqrt{\sum_i (B_i - 4)^2}} \tag{2}$$

この式は、7 段階評価のうち、評価付けの統計的データから真ん中の値である評価値 4 を基準点とし、その値より大きい値はその対象に対して肯定的な評価、小さい値はその対象に対して否定的な評価と分けて相関係数を算出し、これを類似度としている。この場合の基準点は、真ん中の値である評価値 4 を用いている。だが、この手法は、異なる基準点を持つユーザの存在が考慮されておらず、ユーザ間の類似度を算出した場合、正しい評価値の重みが考慮されずに嗜好のズレが生じることが考えれる。

2.4 要求事項

嗜好情報のような情報発信者によってバラつきのある情報は不安定なインデックスであるためにその情報を、検索者の嗜好に合わせて情報提供することは困難とされてきた。しかし、内容ベースフィルタリングや協調フィルタリングの研究から、暗黙的なアクセス履歴を利用したユーザプロファイルの作成や嗜好の定量化、また明示的な嗜好の数値入力からのユーザプロファイル作成や、類似度計算を用いて類似ユーザを発見することで検索者の嗜好を考慮した情報提供を実現しつつある。ユーザが明示的に嗜好を数値入力することは、その人の嗜好を正確に

表す情報を取り扱うこととなり、嗜好に合った情報提供を実現する上で重要である。しかし、既存研究では、評価者がつけたその数値に対する重みが考慮されていないため、そのままの数値から嗜好の類似性を求めた場合、嗜好のズレが生じてしまう。例を用いて説明する。評価者 A がある飲食店 a に対して 3(5 段階評価)という評価付けを行ったとする。A は自身の評価付けの基準として、5 段階評価のうち 3 が「普通・まあまあなの店」、1 が「大嫌い」、5 が「大好き」と考えている。しかし、評価者 B の評価付けの基準は、3 が「好き」、2 が「普通・まあまあの店、1 が「嫌い」の場合、評価者の評価の重み付けによって解釈にズレが生じてしまうことがある。評価値を基にユーザ間の類似度を計算する時、全評価者の評価付けに対する基準点が同じという仮定を用いているが、これは現実的にはありえない。例では 5 段階評価で示したが、これが 7 段階評価や 10 段階評価になると、さらに評価付けの重みによるズレが大きくなり、既存の協調フィルタリング手法が有効に動作しない。したがって、嗜好を考慮した情報提供を実現することは難しいと考える。このような問題を考慮し、それらを解決するため、本研究では以下の機能が必要とされる。

(1) 各ユーザの評価付けの重みを考慮する機能

(2) 評価付けの重みを考慮した評価値を活用し、嗜好の近さを算出する機能

3. 提案手法

本章では、人によってバラつきのある嗜好情報に基準点を設けることで検索者の嗜好に合致する情報を提供する情報推薦システムを提案する。まず、提案システムの概要を述べ、基準点の決定の方法、評価値の重み付け、類似度の計算方法について述べる。

提案システムは、運用するにあたって、ユーザ数の増加といった変化に対する適応性、そして、実験目的であるため実サービスレベルの厳しい推薦時間を要求しないことを考慮した結果、メモリベース法を採用する。

3.1 提案システムの概要

本提案システムは、既存のグルメ検索サイトでは提供されにくい検索者の嗜好に合致する飲食店情報、特に味に対する評価情報を検索者に提供するシステムである。飲食店に対するユーザの評価付けの傾向は人によって異なる。高い評価を多くつける人、評価の甘い人や、低い評価を多くつける人、評価の厳しい人などが考えれる。このような例からわかるように評価値に対する重みは人によって異なる。

本提案システムでは、この評価付けの重みを考慮して、ユーザの嗜好に合致し、高い評価を下すと予測される飲食店情報の収集を行う。そのために他者との飲食店に関する評価の類似度を計算し、類似度の高いユーザからまだ知らない飲食店情報の推薦を受けることができる協調フィルタリングを用いる。本提案システムは、運用するにあたって、ユーザ数の増加といった変化に対する適応性、そして、実験目的であるため実サービスレベルの厳しい推薦時間を要求しないことを考慮した結果、メモリベース法を用いた協調フィルタリングを実現する。

飲食店の推薦を受けるには、類似度の高いユーザを発見する必要がある。そのために嗜好の類似度を算出する必要がある。従来の類似度計算では、計算に用いる評価値の数値をそのまま取り扱っていた。しかし、評価値には評価付けしたユーザ毎に異なる重み付けがされているはずである。そこで、評価値の重み付けの違い、ユーザの基準点を考慮した仕組みを取り入れる。これによって、よりユーザ

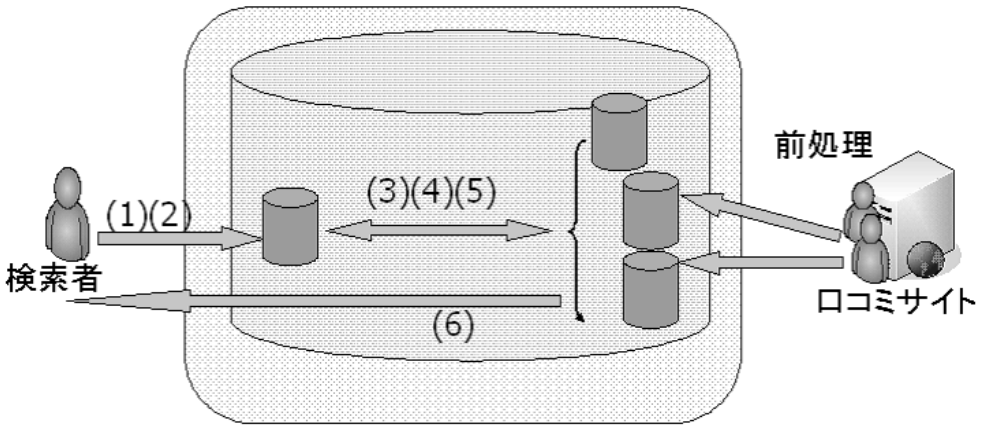

図 3 本提案システムの全体図

の嗜好を反映した類似度を算出することが可能となり、嗜好に合致する情報提供が期待できる。図 3 に提案システムの全体図とシステムの流れを示す。

前処理：推薦されるコンテンツは口コミサイトからコンテンツを収集してくる。収集したコンテンツから飲食店の基本情報、評価情報、ユーザ情報を抽出して利用する。

1. 初めて利用する検索者はユーザ登録を行い、システムにログインする。
2. 検索者は、行ったことのある飲食店の評価値を入力する。
3. 入力された飲食店に評価付けをしているユーザを発見し、評価値を比較する。
4. 評価値比較のとき、そのままの数値ではなく、評価付けの重みを考慮した数値に変換する。
5. 評価付けの重みを考慮して嗜好の類似性の高いユーザを探す。
6. 類似度の高いユーザからオススメ飲食店情報を収集する。

3.1.1 基準点の決定

各ユーザによって嗜好が異なるため飲食店に対しての基準点が異なることを述べてきた。基準点を求めることは、各ユーザの評価付けから、そのユーザにとってどの値が普通・まあまあな評価なのかを選定することを意味する。ここでアンケート収集に関する考えを利用して基準点の選定を行う。ユーザにある対象に対して両極に「好き」、「嫌い」真ん中に「普通・どちらでもない」を設定した評価値を持つアンケートを取ったとき、アンケート結果に真ん中の「普通・どちらでもない」に偏る傾向が見られる。これから、ユーザがその対象に対して特別強い気持ちを持っていない限りは真ん中の「普通・どちらでもない」の値に評価付けしやすいことがわかる。この例からユーザにとって「普通・どちらでもない」評価は最も気軽に評価付けが行える値であり、その値より高い評価ならば「好き」であり、低い評価であれば「嫌い」と分けることができる。この判断を飲食店の評価に利用すると、数段階の評価を用いたアンケートの場合では、最も多く評価した値が、ユーザにとって美味しいわけでもなく美味しくないわけでもない、「普通・どちらでもない」味、つまり基準点となる。よって基準点には、ユーザの飲食店に対する評価値の中で、最も度数の高い（最も評価付けに使用されている）値を基準点に用いる。

ただし、ユーザの評価分布によって、最も高い度数が複数存在する場合がある。そこで、最も高い度数の値が複数存在するような場合、例として「2.5」、「3」、「5」が同等の度数 2 を持つとする。このときの基準点の決定方法は、複数ある最も高い度数の評価値から平均を算出し、これを基準点とする。つまり、((2.5*2)+(3*2)+(5*2)) /6=3.5 より、基準点は 3.5 とする。ただし、算出した基準点が割り切れない値といった理由から評価値の項目 (0.5 から 5 の間の整数) に含まれていなかった場合は、最も近い評価値の項目を基準点にする。または、提案手法によって割り切れなかった値をそのまま利用することも考える。

3.1.2 評価値の重み付け

重み付けに関しては、比較検証のために以下のように、基準点を基に「好き」「普通」「嫌い」の 3 つの区間に分ける方法と基準点を中心として距離を評価付け

の重みと考える方法の 2 種類の手法を提案する。

1. 基準点を中心に「好き」、「普通」、「嫌い」の 3 つの区間に分ける方法

   基準点が割り切れない値の場合は、最も近い評価項目に含まれる数値を基準点とする。例として、算出した基準点が 3.3 の場合、評価項目に含まれない数値のため、最も近い距離にある 3.5 を基準点とする。

   - ユーザ毎に基準点を決定した後、その基準を基に評価値を「好き」、「普通」、「嫌い」の 3 つに分ける
   - 評価値の重み付けに関しては、「好き」＝ 1.5、「普通」＝ 1、「嫌い」＝ 0.5、のようにつける基準点を用いてユーザ A,B の評価値の重み a_n,b_n は以下のように表すことができる。

$$a_n, b_n = \begin{cases} 1.5(M(x) < x_n) \\ 1(M(x) = x_n) \\ 0.5(M(x) > x_n) \end{cases} \tag{3}$$

   ユーザ A,B の飲食店の評価値:$(x_1 \ldots x_n)$

   ユーザ A,B の基準点:$M(x)$

   この式を利用して重み付けを考慮した値に変換する。

2. 基準点を中心として距離を評価付けの重みと考える方法

   この手法では、基準点からの距離が重みとなるため、より重みを再現するために基準点が割り切れない値であった場合でもその値をそのまま基準点として用いる。

   - 基準点が「3.5」の場合、「4」＝+0.5、「4.5」＝+1、「5」＝+1.5「3」＝-0.5、「2.5」＝-1、「2」＝-1.5 のように重み付けを行う。基準点を用いて重み付けを考慮した値に変換する場合以下のように表せる。

$$a_n = x_n - M(x) \tag{4}$$

   ユーザ A の飲食店の評価値:$(x_1 \ldots x_n)$

ユーザ A の基準点:$M(x)$

算出する類似度に負の数の出現なくすために、結果に 5（基準点になり得る最も高い数値）を足す処理を行う。

$$a_n = x_n - M(x) + 5 \tag{5}$$

この方法は、ユーザによって異なる基準点を用い、距離によって評価の重み付けを実現する。同じ数値の比較であっても基準点からの距離の違いで、異なる重み付けとなりユーザの飲食店に対する評価の重みをより再現できる。

3.1.3 類似度計算

ユーザ毎に飲食店の評価値に対する重み付けを行った後、評価値の重みをベクトル要素としてベクトル空間モデルを用いて類似度 $\cos\theta$ を算出する。類似度計算には一般的にポアソン相関係数と Cosine Similarity が協調フィルタリングで広く用いられているが、互いに両式に変換が可能なため、今回はシンプルな Cosine Similarity を用いて類似度計算を行う。

$$cos\theta = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}|\,|\vec{b}|} \tag{6}$$

$\vec{a}$,$\vec{b}$:ユーザ A,B の飲食店に対する評価値の重みベクトル

3.1.4 コンテンツの収集

協調フィルタリングを用いて情報推薦を行うには、ユーザが評価付けを行い、その評価を基に推薦されるべきコンテンツが必要である。本システムでは、飲食店情報を含んでいるコンテンツを対象とする。飲食は人間の生活にとって、欠かすことのできない行為であり、この行為を日常生活の出来事として口コミサイトへ投稿、個人ブログで情報公開している人も多い。これらの中から評価を含んだ飲食店情報を掲載しているサイトをコンテンツとして収集する。飲食店情報に関しては、個人ブログから様々な視点で記述された評価情報を閲覧することができ

る。また、複数の評価情報を基に客観的にその飲食店を評価することもできる。しかし、個人ブログは評価情報を含んでいる可能性は高いが、必ずしも飲食店に関する情報を含んでいるとは限らない。また、公開情報からある飲食店に対する評価情報を取り出すにも、テキストから情報発信者の明確な嗜好を抽出して数値化することは困難である。そのため、あらかじめ飲食店に関する評価情報を数値として入力している情報を収集することで、協調フィルタリングに利用する。

価格.com が提供しているサービスに食べログ.com（以後、食べログ）がある。食べログでは、ユーザはまずユーザ情報（メールアドレス、住所、性別など）登録を行う。食べログには、登録ユーザが飲食店に対して入力した評価値やコメントを蓄積している DB がある。この DB 内の情報から、新規ユーザは自身の行ったことのある飲食店を検索して評価値（10 段階評価）やコメントを入力する。図 4 にユーザが登録した飲食店の評価値やコメントが表示されたページの例を示す。図 4 の最上部に、登録ユーザのユーザ名（魔人ブウ）、中段左に飲食店の店舗名 (らすた)、中段中央に飲食店に対する評価値、最下部にコメント文のようなページ構成になっている。

また図 5 は、飲食店の基本的な情報が含まれる飲食店トップページの例である。最上部を見ると飲食店の店舗名、住所、電話番号、休日などの基本的な情報が含まれている。中段から最下部にかけてその飲食店を評価したユーザの評価情報が表示されている (☆の数が評価値となっている)。食べログは、飲食店の情報を閲覧する・公開するという目的で利用するサイトであるので、飲食店の評価情報を収集しやすい。また評価情報が数値 (図 4 中段中央) として入力されており、情報発信者の嗜好を抽出しやすいため、嗜好を数値化する変換をする必要がない。よって、評価付けの重みを考慮した推薦システムを構築する上で食べログの所有するコンテンツは適していると考えられる。

図 4 「食べログ」ユーザトップページ

図 5 「食べログ」飲食店トップページ

4. 情報推薦システムの設計

本章では、前章の提案をうけて、評価付けの重みを考慮した情報推薦システムの設計を行う。まずシステムの設計概要を述べ、システムの詳細設計を述べる。

4.1 システムの設計概要

提案システムでは、まず、利用者は新規登録ページでユーザ情報であるユーザ名とパスワードを DB(Data Base) に登録する。次にトップページにて、登録したユーザ名とパスワードを入力すると提案システムにログインすることができる。

ログイン後、ユーザの個人ページへ遷移する。個人ページでは、都道府県を指定して飲食店の店舗名を検索、または店舗名を入力することで評価付けする飲食店の検索を行う。評価付けする飲食店を検索した後、10 段階評価を使用して味に関しての評価付けを行う。評価付けを行うと、評価したユーザ名、評価値が DB へ登録される。以後ログインする度にユーザが過去に登録した店舗名とその店に対する評価値が出力される。過去に評価した飲食店の評価値を出力させることで、新しく評価する飲食店は過去に評価した飲食店の評価値考慮しながら評価付けが行える。評価付けを 5 つ以上行った後、検索する場所を指定してオススメ検索を行う。このときオススメ飲食店に関する情報は、DB 上に登録されている嗜好の近いユーザから得ることになる。嗜好の近さは 3.1.2、3.1.3 の方法を用いて、ユーザの評価値とその評価に対する重み付けを他ユーザと比較することで算出する。算出した嗜好の近さを基に、他ユーザの評価している飲食店でユーザが評価しておらず（まだ知らない)、高い評価が予想される飲食店情報を出力させる。

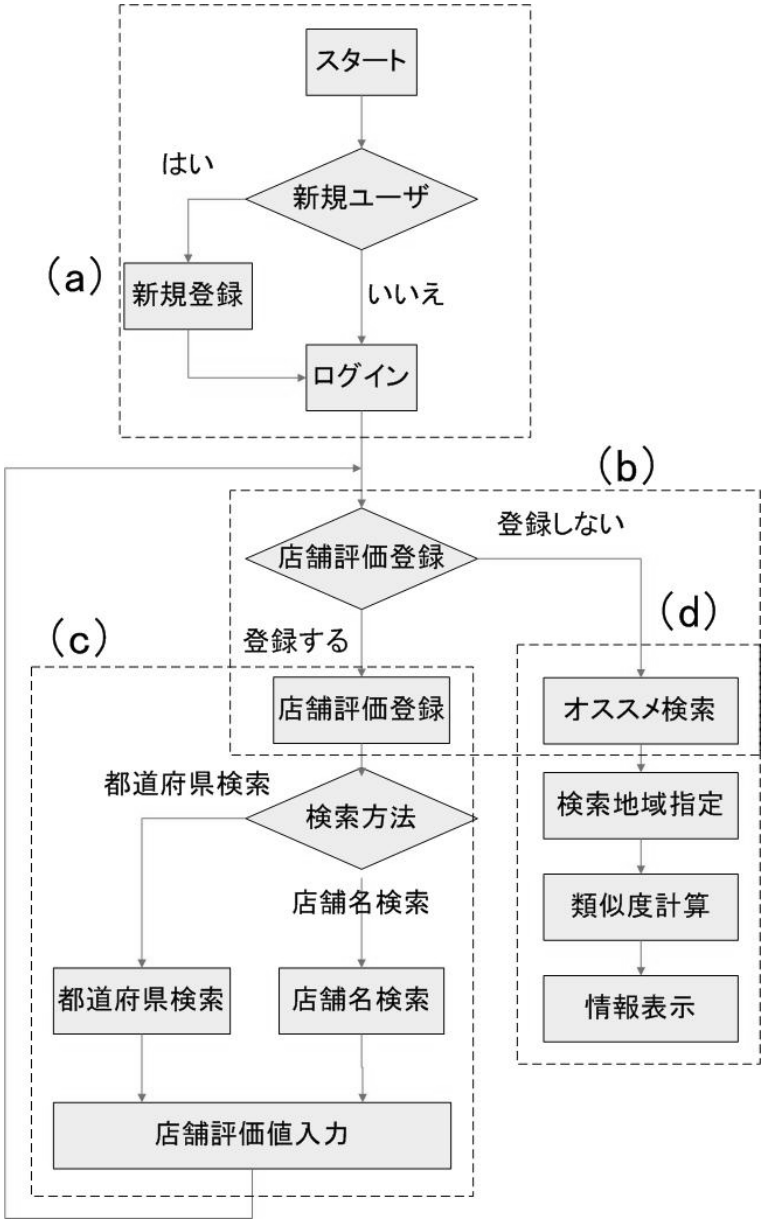

図 6 提案システムのフローチャート

提案システムの動作を図 6 に示す。

- 新規登録処理 (図 6(a))
  - 利用者ユーザはユーザ名とパスワードをシステムの DB へ登録。
- ログイン処理 (図 6(b))
  - 利用者ユーザは登録したユーザ名とパスワードを入力することでシステムへログインする。
  - ログイン後、新規ユーザは店舗に対して評価を行うため店舗評価登録処理へ移る。店舗評価登録が住んでいるユーザは、さらに店舗評価登録を行うか、嗜好に合った飲食店を探すためのオススメ検索処理へ移る。
- 店舗評価登録処理 (図 6(c))
  - 店舗評価登録処理では、店舗評価を対象となる店舗を検索するために、都道府県検索と店舗名検索が行われる。
  - 検索結果として店舗名のリストが返され、リスト内から評価付けをする店舗名を選定。
  - 店舗名を選定した後、評価値を入力する。
  - 評価値入力後、店舗評価登録処理を繰り返すか、嗜好に合った飲食店を探すためのオススメ検索処理へ移る。
- オススメ検索処理 (図 6(d))
  - 指定した地域の飲食店情報を得るために都道府県、そしてさらに詳細な範囲指定である都市を選択する
  - 利用者ユーザの評価情報を基に、共通の飲食店に評価付けしているユーザと飲食店に対する評価値から類似度を算出する。
  - 高い類似度を持つ類似ユーザの飲食店情報の中から、指定した地域に属し、類似ユーザの基準点よりも高い評価を受けている飲食店情報を表示する

4.2 システムの詳細設計

4.2.1 コンテンツの収集方法

コンテンツの収集に関しては、情報量の変化（新規登録された飲食店や新規登録ユーザ）、また情報の鮮度を考慮すると実運用上、定期的に収集することが望ましい。しかし、本研究ではユーザの評価付けの重みに着目しているため、あらかじめ食べログから十分な量のコンテンツを収集して DB に格納したものを利用する。

2.2.1 や 2.2.2 で述べたように、嗜好に合った情報提供を実現するためには、安定したインデックスである飲食店に関する基本情報と不安定なインデックスである飲食店に関する評価情報が必要である。そのために、必要な情報を含む食べログの各ユーザのトップページと各飲食店のトップページの HTML ファイルを収集して、情報を抽出する。

- ユーザトップページと飲食店トップページの収集

  飲食店の基本情報と評価情報を抽出するために、URL を指定して HTTP などを用いて Web からファイルをダウンロードしてくるツールである wget [7]を用いる。これを用いて、必要情報が含まれている食べログの各ユーザのトップページと各飲食店のトップページの HTML ファイルを収集する。

- 必要情報の抽出

  収集したユーザトップページと飲食店トップページの HTML ファイルは構造化されている。そのため HTML タグ構造を解析し正規表現による文字列パターンマッチングを用いて必要情報を抽出する。以下に例を示す。

  – 飲食店の店舗名・住所などの飲食店情報（飲食店トップページから抽出）

    ＜ li class=”name”＞＜ strong ＞店舗名＜strong ＞のように class の属性値に name が使われており、店舗名の抽出が容易な構造になっている。よって＜ strong ＞ (.+?) ＜ strong ＞のような正規表現による文字

[7]GNU Wget:http://www.gnu.org/software/wget/

列パターンマッチングを用いることで抽出できる。同様に住所の場合＜li class=”address”＞(.+?)＜strong＞で抽出できる。

– ユーザの飲食店に対する評価値（ユーザトップページから抽出）誰がどの飲食店に評価付けしているかの抽出。

また評価付けの対象となる飲食店の店舗名と評価値に関しては、＜IMG class=title_ico alt=料理・味 src=”/images/ico_rating_food.gif”＞＜IMG alt=”” src=”/images/bar_dtl_40.gif”＞のようにタグ内に含まれるalt属性から「料理・味」に対しての評価であると確認でき、「～/bar_dtl_40.gif」の箇所から評価値を抽出できる。よって「料理・味」の文字列が出現した後に＜/bar_dtl_(.+?).gif＞により抽出する。

以上の方法により、飲食店の店舗名、住所、都道府県、電話番号、休日、営業時間、飲食店HPのURL、ユーザ名、ユーザの飲食店に対する評価値、文章を抽出できる。

4.2.2 DBへ格納

4.2.1の方法で抽出した情報をDBに格納するためのテーブルを作成する。

システム利用前のDBには、食べログから抽出したユーザ情報、飲食店情報、ユーザの飲食店に対する評価情報を格納する。表2は、飲食店トップページ(図5)から抽出した飲食店情報を格納するテーブルである。カラムに含まれる店舗名、店舗の都道府県は店舗評価登録処理を行うために必要な情報である。これらの情報は、店舗名検索を行い店舗を選定するとき、またその店舗名を検索するために店舗の都道府県・都市で絞込みをかけるときに必要な情報である。これによりユーザの飲食店の評価のために必要な店舗名を探す行為を負担なく行える。

表3は、ユーザトップページ(図4)から抽出した食べログに登録されているユーザ情報と初めてのシステム利用で発生する新規登録のユーザ情報を格納するテーブルである。提案システムを利用するユーザは、新規登録としてユーザ名、パスワードをDBに格納し、これを利用してトップページであるログイン画面からログインする。

表 2 店舗情報テーブル (rst)

項目	データ型	説明
rst_id	int	店舗 ID
rst_name	char(30)	店舗名
area1	char(10)	店舗の都道府県
area2	char(30)	店舗の都市
address	text	住所
tel	char(30)	電話番号
holiday	char(30)	定休日
hour	text	営業時間
url	text	店舗 HP の URL

表 3 ユーザテーブル (user)

項目	データ型	説明
user_id	int	ユーザ ID
user_name	char(30)	ユーザ名
passwd	text	パスワード

表 4 評価テーブル (rst_value)

項目	データ型	説明
id	int	識別 ID
user_id	int	ユーザ ID
rst_id	int	店舗 ID
comment	text	飲食店へのコメント文
value	float	評価値

表 5 ユーザの基準点テーブル (criterion)

項目	データ型	説明
user_id	int	ユーザ ID
criterion1	float	基準点 1
criterion2	float	基準点 2

表 4 は、ユーザトップページ (図 4) から抽出した食べログに登録されているユーザが評価付けを行った店舗に対しての評価値とコメントを格納するテーブルである。また利用者ユーザが評価付けを行った店舗の評価値も格納されるテーブルである。

ログイン後、利用者ユーザは嗜好の類似するユーザを探すために自身の嗜好情報として飲食店に対する評価値の入力、また登録ユーザとの飲食店に対する評価値の比較を行う。

表 5 は、3.1.1 で決定したユーザ毎の基準点を格納するテーブルである。このテーブルにはユーザが評価付けを行った評価値を基に基準点を格納する。そのため、利用者ユーザが評価付けを行う度に基準点を更新する。テーブル内に基準点のカラムが 2 つあることに関しては、2 つの提案手法で基準点が異なることが起こるためである。

3.1.2 で述べた、基準点を中心に 3 つの区間に分ける手法では、基準点が評価項目に含まれるような値になる。また、基準点を中心に距離で重み付けをする手法では、割り切れない小数値をそのまま基準点として用いる。よって 2 つの手法を実験するために格納するカラムを 2 つ用意している。

4.2.3 類似度計算

3.1.3 で述べた類似度計算の方法により、DB 内に登録されている評価情報を基にユーザ間の類似度を計算する。類似度計算には、評価値と基準点を用いて算出した評価付けの重みを用いる。

まずシステムは、システム利用者であるユーザが DB 内に登録した飲食店と同

じ飲食店（以後共通店舗）に評価付けをしているユーザの発見を行う。共通店舗を持つユーザごとに、評価テーブルと基準点テーブルを用いて評価付けの重みを算出する。システム利用者であるユーザと類似ユーザの評価付けの重みを、Cosine Similarity の式に代入することでユーザ間の類似度を測ることができる。

4.2.4 推薦結果表示

3.1.3 で述べた類似度の高い類似ユーザから、ユーザが未開拓で高い評価を下すと予想される飲食店の基本情報と評価情報を推薦結果として表示する。推薦結果は類似度の高いユーザ順に表示され、各ユーザ毎に類似ユーザの基準点よりも高い評価付けされている飲食店に関する情報を推薦する。

5. 情報推薦システムの実装

本章では、前章の設計に基づいて実装を行った開発環境と、その詳細について述べる。

5.1 開発環境

実装した機能は、ユーザ毎の評価値の重みを考慮するための基準点の選定機能、および類似度算出機能である。これらを Web サーバ上の CGI(Common Gateway Interface) として稼動させた。開発環境の詳細を表 6 に示す。開発に使用した言語は PHP である。またデータベースには MySQL を使用した。

5.2 利用画面例

提案システムの処理手順は、図 6 で示すとおりである。まず、トップページであるログイン画面から新規登録処理を行い、新規登録後ログイン処理に移る。ログイン処理後の利用画面は、図 7 のようになっている。2 つのフレームで構成されており、上のフレームは、ユーザ名の表示、また評価登録を行うための店舗検索機能を提供している。下のフレームは、ユーザが登録した飲食店の店舗名と評価値を順に表示する。また、場所を指定した検索を可能にするため、都道府県を指定するためのプルダウンメニューを設置している。次にユーザが店舗評価登録を

表 6 開発環境の詳細

CPU	AMD Opteron252*2 2.6GHz
Main Memory	2GB DDR/400
OS	Fedora Core5
Web サーバアプリケーション	Apache/2.2.2
実装言語	PHP5.1.6
データベース	MySQL5.0.22

図 7 店舗評価登録前

行った画面を図 8 に示す。上のフレーム内にある都道府県検索ボックスを利用して評価付けする飲食店を検索する。大阪にある飲食店に対して評価付けを行う場合、都道府県検索ボックスに「大阪」と入力し決定すると、プルダウンメニューが出現し、大阪にある飲食店リストが閲覧可能となる。評価付けをする飲食店を決定した後、評価値を選択するプルダウンメニューから評価値を決定し、「店舗評価登録」ボタンを押す。下のフレームには店舗評価登録をした店舗名と評価値が順に表示される。ユーザは、登録処理を繰り返すことで嗜好を表す評価情報を格納する。ある程度評価情報を入力した後、検索したい都道府県、都市を場所指定のプルダウンメニューから選択し、「オススメ検索」ボタンを押すとオススメ検索処理へ移る。

図 9 に、ユーザがオススメ検索をした検索結果を示す。下のフレームにオススメ検索の検索結果が表示されている。フレームの一番上に表示されているのは、最も類似度の高いユーザのユーザ名とその類似度である。それ以降には、そのユーザがシステムに登録した登録店舗数 (経験店舗数)、類似度計算に用いた共通店舗

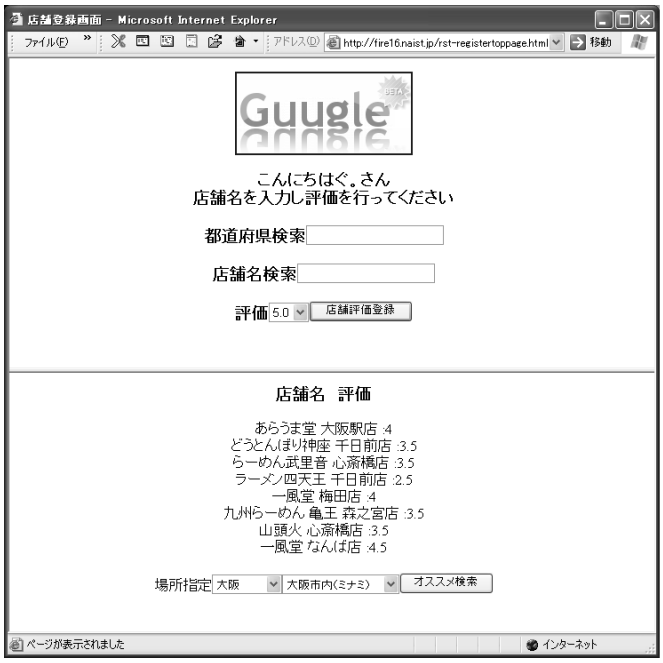

図 8 店舗評価登録後

の数 (比較店舗数)、類似ユーザの基準点、評価平均、そしてオススメの店舗名と評価値、飲食店の基本情報 (住所、電話番号、休日、営業時間、店舗 HP)、飲食店に対するユーザの一言が表示されている。オススメされる飲食店は、類似度の高いユーザの基準点よりも高い評価付けをされている飲食店が評価値の高い順に表示される。また上のフレームは利用者ユーザが、さらに店舗評価登録を行えるように表示させているものである。

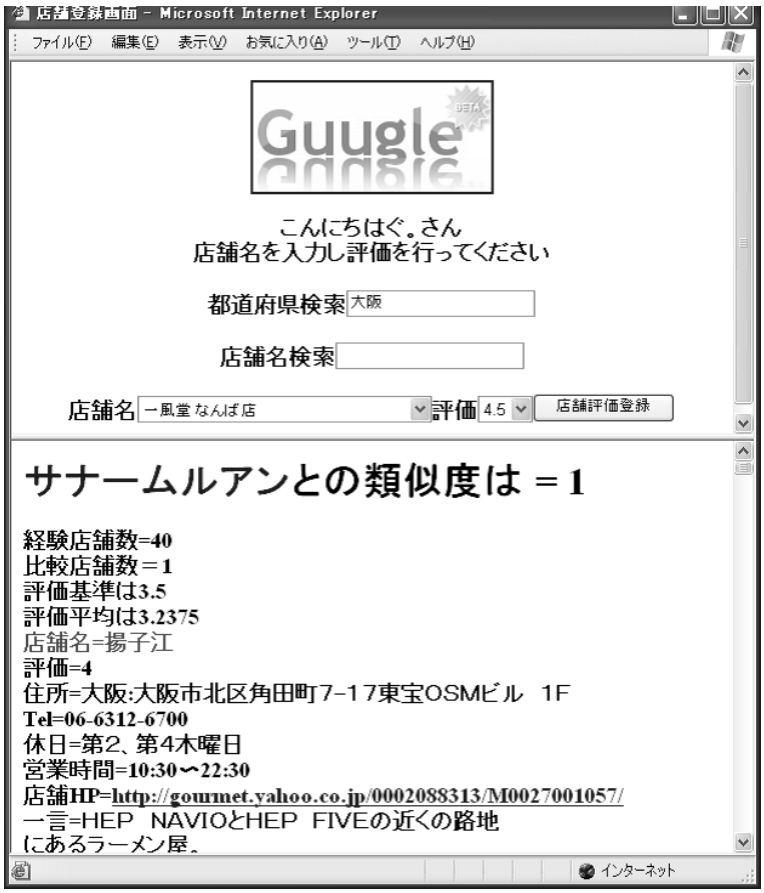

図 9 オススメ検索結果の表示画面

6. 評価

本章では、本研究の提案システムが嗜好に合った情報提供を実現できているかの評価を行う。異なる視点で提案するシステムを評価するために3つの実験を行った。基準値同定の性能評価の結果を 6.2 に、実証実験を行い類似ユーザの評価結果を 6.3、実証実験を行い推薦店舗の評価結果を 6.4 に示す。

6.1 評価環境

表 7 評価環境

DB 内の登録ユーザ数	1022 人
登録店舗数	4216 件
一人当たりの評価付け数	約 10 件
総評価付け数	11066 件

表 8 各実験の被験者数

実験 1（基準点同定の性能評価）	100/1022 人
実験 2 （実証実験：類似ユーザの評価）	6 人
実験 3 （実証実験：推薦店舗の評価）	9 人

実験 1 では、既存ユーザ 100 人のサンプリングによる性能評価を行った。実験 2 では、研究室内学生 6 人、実験 3 では、研究室内学生 4 人、他研究室学生 1 人、他研究科学生 2 人、学内事務員 2 人の計 9 人に提案システムの性能評価に協力してもらった。

6.2 実験 1：基準点同定の性能評価

提案システムは、ユーザの異なる嗜好に対して基準点を設定することでユーザ間の嗜好のズレをなくす機能を実現し、嗜好に合った飲食店情報を提示する。比

表 9 検証結果例

登録ユーザ	基準点→評価値	オススメ店舗	推薦者ユーザ	推薦者の基準点→評価値
ユーザ 1	4 → (3.5:隠す)	店舗 A	ユーザ A	4 → 4
			ユーザ B	4 → 4
			ユーザ C	4.5 → 4.5
			ユーザ D	3 → 3
			ユーザ E	3.5 → 4.5
誤り数の合計				1

較には、基準点にユーザ間の共通店舗の評価平均を利用した場合 [1]、ユーザの評価のスケールに対して中央に位置する値を基準点と利用した場合 [6] を用いる。提案システムを評価するため、システム内のデータベースからユーザを取り出し、ある評価値を隠した後に新規ユーザとして登録させ、そのユーザへの推薦内容を検証する。以下、手順を示す。

1. 数名のユーザを DB からランダムに抽出する。
2. 抽出した各ユーザの店舗評価情報のバックアップをとっておき、そのユーザに関する情報を DB から消去する。
3. 抽出したユーザの中から 1 人選択し、記録した店舗評価情報を基に、評価値の低い 1 つの店舗以外の店舗情報を登録する (評価値の低い店舗評価を隠す)。
4. 登録が終わったらオススメ検索を実行し、評価付けを隠した店舗の推薦の有無を確認する。

このようにシステム性能評価には、推薦されるべきものではない対象の数である誤り数を計上する。推薦結果として出力される情報は、類似ユーザの基準点よりも高い評価値を持つものである。

表 10 異なる基準点毎の誤り数

	提案	平均値	基準点 3	基準点 3.5	基準点 4
誤り数	23	43	88	67	15

表 9 は、ユーザ 1 を新規登録し店舗評価登録をした後のシステムの推薦結果である。ユーザ 1 は、店舗 A に対して評価付けを行っているが、推薦システムの検証のために隠す。システムがユーザ 1 が「嫌い」(基準点より低い) と評価付けしている店舗 A を推薦しないことを確認する。基準点が 4 であるユーザ A は店舗 A に対して 4 の評価付けをしている。これはユーザ A にとって店舗 A は「普通・まあまあ」な評価となる。ユーザ B、C、D も同様である。この場合類似ユーザである推薦者の基準点よりも低い数値のため、システムはユーザ A、B、C、D からユーザ 1 への推薦を行わない。3.5 の基準点を持つユーザ E による店舗 A に対して、4.5 の評価付け (「好き」) をしている。よってシステムはユーザ 1 にユーザ E による店舗 A に対する情報をオススメ情報として提示する。ユーザ 1 は店舗 A に対して「嫌い」の評価付けを行うことが既知であるため、ユーザ E のオススメ情報はユーザ 1 にとっては誤りになる。よってユーザ 1 に対しての誤り数は、ユーザ E からユーザ 1 への推薦である 1 つになる。

このような検証をランダムに抽出した 100 人のユーザに対して基準点毎に行った。基準点には、提案するユーザの評価付けの傾向から最も度数の高い評価値を基準点とする方法、ユーザ間の類似度比較に用いられた共通店舗の評価値の平均値を用いる方法、評価のスケールの中央の値を採る方法 (基準点 3,3.5,4) の 3 つの方法で同定した点を用いる。ユーザにこれらの基準点を用いた場合、誤り数が最も少ない方法を有効な基準点としてカウントする。基準点毎にカウントした結果を表 10 に示す。

抽出した 100 人のユーザを手順に示すとおりに登録し、これより提案する方法が既存の基準点同定方法 (平均値、評価のスケールの中央値) よりも誤り数の少なく有効であることがわかった。しかし、表 10 からもわかるように基準点をより高い値に同定することで誤った推薦 (検索者が「嫌い」だと考える店舗推薦) を行わないことがわかる。高い数値に基準点を同定することで、より高い評価付けを

表 11　異なる基準点毎の情報推薦数

	提案	平均値	基準点 3	基準点 3.5	基準点 4
推薦数	23	8	77	0	0

受けている店舗しか推薦されない。よって、検索者が高い数値を基準点に同定していない限りは、「嫌い」な店舗推薦を受けることはない。この結果より本研究で用いた食べログの DB 内に含まれるユーザの評価付け傾向が比較的に高い数値に収まっていることがわかる。

また、表 11 は、異なる基準点毎の情報推薦数を表している。表 11 より、「4」のように基準点が比較的高い場合、推薦数が 0(ゼロ) になっている。これは、基準点を「4」と固定した場合、「4」よりも高い評価付けを行ったことのないユーザは、推薦する情報がなくなるためである。よって比較的高い値に基準点を固定した場合、全体の推薦数が減少する。さらに都道府県や都市で情報を絞り込むため、推薦情報が減少、または 0(ゼロ) になる。この結果より基準点を固定するより、提案する基準点同定方法のようにユーザの評価付けに合わせた基準点を同定する必要がある。

6.3 実験 2(実証実験)：類似ユーザの評価

類似ユーザの評価では、提案システムが発見した類似ユーザが検索者の嗜好に合っているかの確認をする。実証実験には、被験者として研究室内学生 6 人に協力してもらった。各被験者が行う手順を以下に示す。

1. システムにログインする。
2. 行ったことのある店舗の評価登録をする。
3. 特定の場所で (実験 2 では「大阪」または「京都」) 検索を行う。
4. 得られた検索結果から類似度の高いユーザを上位 5 人選ぶ。

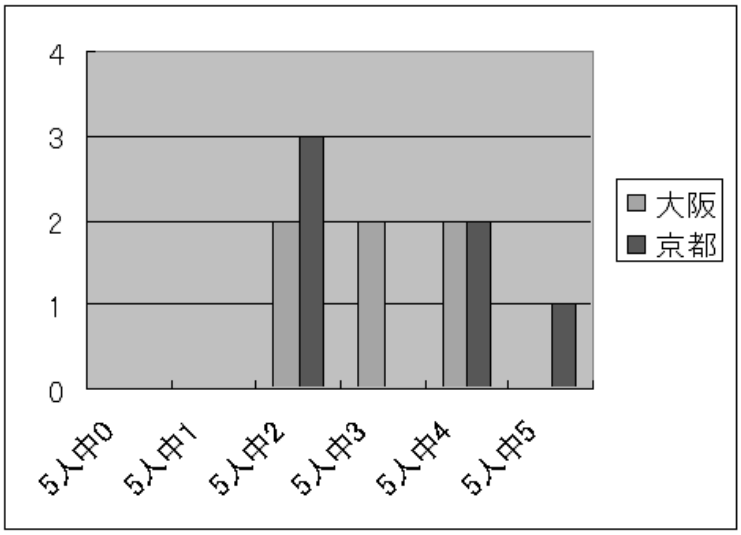

図 10 嗜好に合ったユーザの割合

5. 選んだユーザと嗜好が似ているかを確認する。

   食べログからユーザ情報を確認し、評価した店舗に対しての文章、評価値を見ることで主観的に嗜好が似ているかを判断する。

6. 類似度が高いユーザ 5 人の中に嗜好が似ていると思えたユーザが何人いたかをカウントする。

実験 2 の結果をアンケート集計してまとめたものを図 10 に示す。アンケートには、「店舗情報を推薦するユーザが嗜好に合ったユーザであったか？」という内容を用いた。図 10 から、提案システムが類似ユーザとして表示したユーザが被験者の嗜好にどの程度合っていたかがわかる。研究室内学生 6 人がそれぞれ検索を行った結果、嗜好に合っているユーザが上位 5 人の中に 1 人もいない、または 1 人しかいないと答えた人が 1 人もいなかった。提案システムにより表示された類似ユーザの上位 5 人中少なくとも 2 人以上嗜好に合ったユーザがいることがわかった。よって、被験者の嗜好に合ったユーザを多く提示できたといえる。

6.4 実験 3(実証実験)：推薦店舗の評価

推薦店舗の評価では、実証実験を踏まえて提案システムが検索者の嗜好に合った情報提供を実現できているかの確認する。実証実験には、被験者として研究室内学生、他研究科学生、学内事務員の計 9 人に協力してもらった。各被験者が行う手順を以下に示す。

1. ユーザ名、パスワードを登録し、その情報を用いてシステムにログインする。
2. ログイン後、行ったことのある店舗の評価登録を数店舗する。
3. 検索場所を「大阪」にし、さらに細かい範囲の都市指定では任意の場所を選び、オススメ検索を行う。
4. 検索結果で得られた推薦店舗情報の中から数店舗選び、実際に足を運ぶ。
5. 実際に足を運んだ結果、その店舗が被験者の嗜好に合っていたかどうか確認するため、アンケートに回答してもらう。
6. この手順を検索場所「大阪」と「京都」で 1 回ずつ行う。

実験 3 の結果をアンケート集計してまとめたものを図 11 に示す。アンケートには、「推薦店舗は嗜好に合っていたか？」という内容を用いた。

図 11 はアンケートの集計結果を表しており、提案システムが嗜好に合った店舗情報の提供を実現できたかがわかる。被験者 9 人に提案システムを利用してもらい、「大阪」と「京都」で検索を行ってもらった。その結果、嗜好に「合っていた」と思える店舗が 11 件、嗜好に「やや合っていた」と思える店舗が 3 件推薦された。よって、概ね被験者の嗜好に合った店舗情報を提供できたといえる。

6.5 今後の課題

本提案システムでは、ユーザ毎に基準点を設定することで嗜好に合った情報提供を実現した。しかし、ユーザによって評価分布に特徴があり、全てのユーザに対して適切な基準点を設定できたとは言い切れない。今後は、ユーザの評価分布

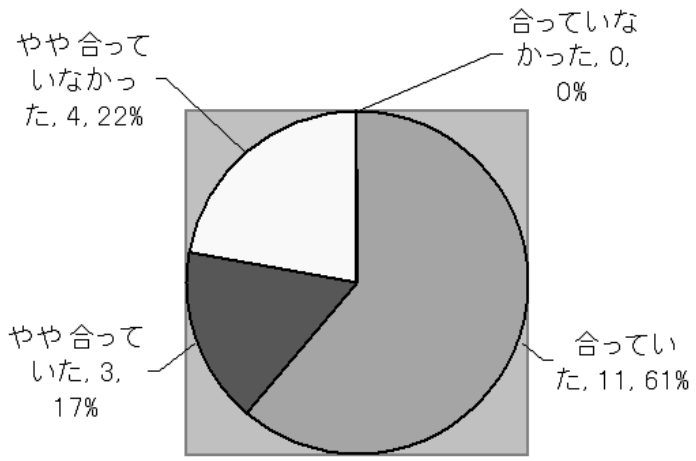

図 11　推薦店舗に対する被験者の嗜好の一致性

をさらに細分化することでシステムの性能向上を図る。また、類似度を算出する際に評価値や評価値の重みだけでなく、評価付けしている対象の性質・特徴を分析して、対象間の類似度を算出することでより嗜好に合った情報収集を可能にする。実証実験に協力してくれた被験者から、その日の天気、被験者の気分、時間帯などその時の状況に合わせた情報提供を実現することも重要であるという意見を頂いた。つまり評価付けの重みを考慮した情報推薦だけではなく、コンテキストを考慮した情報推薦を実現することが嗜好に合った情報提供を実現するには必要である。

7. おわりに

多くのユーザは音楽 (TSUTAYA Onlin) やグルメ (ぐるなび、ぐるめぴあ) などの目的に特化した検索サービスを利用することで嗜好に合った情報を収集してきた。このようなサービスは、事業主が広告として情報を掲載するよう依頼するため、情報としての信頼性が高く検索者は安心して見れる。しかし、発信される情報は事業主側に不利益にならないような情報しか公開せず、検索者が本当に知りたいその対象の良し悪しを判断するための評価・評判情報が含まれていない。そこで個人が SNS やブログなどに含まれるある対象への評価情報が注目されている。検索者はある対象への評判情報を収集したい場合、複数の個人サイトや口コミサイトを利用して記述されている内容を吟味し自身の嗜好に合うかの予想をする。評価対象に対して自身と同等の評価を下している人は嗜好が似ている人だと考え、嗜好の似ている人の評価から検索者の未知のものに対しての評価を予測する。このとき検索者は嗜好の類似を記述文章から判断しているが、実際に機械処理によって記述されている内容から情報発信者の嗜好を抽出することは人によって嗜好表現も異なり、またその度合いも異なるため困難である。そこで嗜好情報を検索者の行動や明示的な入力により数値化し、その値を利用することで嗜好に合った情報を提供する協調フィルタリングの研究が行われてきた。数値化された嗜好情報から嗜好の類似度を算出し、類似度の高いユーザから検索者の未知の情報を推薦する。このとき、類似度算出に用いられる数値に評価付けをした評価値の正しい重みが考慮されていない。

そこで本研究では、人の嗜好に基準点を設定することで、検索者と情報発信者との嗜好のズレをなくし、嗜好に合った情報提供システムを提案した。提案したシステムを設計、実装し数名の学生に利用してもらった結果から、基準点をユーザ毎に設定することで既存技術よりも嗜好に合った情報提供ができていることを確認した。本提案システムでは、ユーザ毎に基準点を設定することで嗜好に合った情報提供を実現した。しかし、ユーザによって評価分布に特徴があり、全てのユーザに対して適切な基準点を設定できたとは言い切れない。今後は、ユーザの評価分布をさらに細分化することでシステムの性能向上を図る。また、類似度を算出する際に評価値や評価値の重みだけでなく、評価付けしている対象の性質・

特徴を分析して、対象間の類似度を算出することでより嗜好に合った情報収集を可能にすること、検索者の気分、その日の天気といったコンテキストを考慮した情報推薦を組み込むことを今後の課題とする。

謝辞

適切な研究指導と豊富な研究環境を提供頂き、また暖かい励ましやご指摘を頂いた、指導教官である奈良先端科学技術大学院大学インターネット・アーキテクチャ講座の砂原秀樹教授およびインターネット工学講座の山口英教授に心より感謝の意を表します。また、研究方針や考え方、論文執筆、設計や実装の詳細にわたる検討など研究全般においてご教示、ご指導頂きました、奈良先端科学技術大学院大学インターネット・アーキテクチャ講座の藤川和利助教授に心より感謝致します。研究の進め方、研究に対する悩みを真剣に受け止めて親身になって支えてくれた奈良先端科学技術大学院大学インターネット・アーキテクチャ講座島田秀輝助手に心より感謝致します。研究に対する取り組み方や、進め方・論文執筆、人としての生き方、研究会での朝方まで続いた発表資料チェックなど非常に熱心で事細かな指導を頂いた奈良先端科学技術大学院大学インターネット・アーキテクチャ講座新井イスマイル氏に心より感謝いたします。

研究活動と学校生活全般において、温かい眼差しで見守っていただき、いつも優しい笑顔と厳しい激を提供してくださった奈良先端科学技術大学院大学インターネット・アーキテクチャ講座の呂悠妃女史に心より感謝いたします。

自身の研究活動に忙しいにも関わらず、システム実装に的確なアドバイスと共に協力をしてくれた佐藤貴彦氏そして、実証実験に協力してくれたインターネット・アーキテクチャ講座の M1 のみなさまに心より感謝致します。また研究室内で運営している電子商店 LOWSAN には、生きるために欠かせない食料、主にインスタントラーメン、駄菓子を安価に提供していただき LOWSAN 運営陣一同に心より感謝いたします

また大学院生活の始まりから終わりまで、600 円という安い値段で最高に美味しいラーメンを提供してしてれたラーメン研究所あまのじゃくの店長とその奥さんに心より感謝いたします。

本研究に取り組むにあたり、遠方にもかかわらず心の安らぎを与えてくださった久留米大学病院の小塩美佳女史、毎日の私生活で心の支えとなってくださった山田愛弥女史に心より感謝いたします。私生活で心配をかけ迷惑をかけたにも関わらず、暖かい眼差しで見守っていただき、将来について真剣に話し合ってくれ

た横地美里女史に心より感謝いたします。

最後に、家族である父川口 雅稔、母博子、兄高大、妹美咲、叔母洋子、祖母秀、愛犬チコ、フクは私を温かく見守り、経済的、精神的な面において支え続けてくれました。そして、友人・知人に心より感謝いたします。

参考文献

[1] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J. ”GroupLens: An Open Architecture for Collaborative Filtering of Netnews”. CSCW ’94 Procedings. pp.175-186. 1994

[2] Morita, M. and Shinoda, Y. ”Information Filtering Based on User Behavior Analysis and Best Match Text Retrieval”. in Proc. of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information retrieval. pp.272-281. 1994

[3] Sakagami, H. and Kamba, T. ”Learning Personal Preference on Online Newspaper Articles from User Behaviors”. Proc. of the Sixth International World Wide Web Conference, In Computer Networks and ISDN Systems, Vol. 29,pp.1447-1456. 1997

[4] 吉田 将志, 吉高 淳夫, ”Digital Reminder:ユーザの視点からの実世界指向データベースの構築とそのインタフェース”, in Proc. of the 8th Workshop on Interactive Systems an Software (WISS’2000), pp.101-110. 2000

[5] 赤木 法生, 大島 裕明, 小山 聡, 田島 敬史, 田中 克己, ”レビューページ例からの属性抽出に基づくレビューページ検索”, DEWS2006, 2C-i10. 2006

[6] Shardanand, U. and Pattie Maes ”Social Information Filtering: Algorithm for Automationg ”Word of Mouth””. CHI ’95 Conference Procedings.pp.210-217. 1995

[7] 神嶌 敏弘 ”情報システムー情報過多時代をのりきる” 情報の科学と技術 情報科学技術協会 情報科学技術公開.Vol.56 No.1,p.452-457 2006