Two-Phase Decision Fusion Based On User Preference [*]

Yi-Shin Chen and Cyrus Shahabi

Integrated Media Systems Center
University of Southern California
Los Angeles, CA 90089-2564, U.S.A.
email: [yishinc,shahabi]@cs.usc.edu

Gully A.P.C. Burns

K-Mechanics Research Group
University of Southern California
Los Angeles, CA 90089-2564, U.S.A.
email: gully@usc.edu

Abstract

In order to prevent information overload on the Web, people sometimes rely on critics' evaluations to narrow down their search for favorite objects. However, due to the subjectivity of humans' perception, the set of favorite objects varies from one person to the other. This is because each object can be reviewed using different criteria, and different people have different reliance on the same set of criteria and/or critics. Therefore, the task of fusing the decisions of different critics considering the users' preferences is challenging. To address this challenge, we introduce a new concept, termed two-phase decision fusion, *where both critics and criteria evaluations are considered in order to personalize a search. In each phase, we utilize various decision fusion operators and investigate their interplay.*

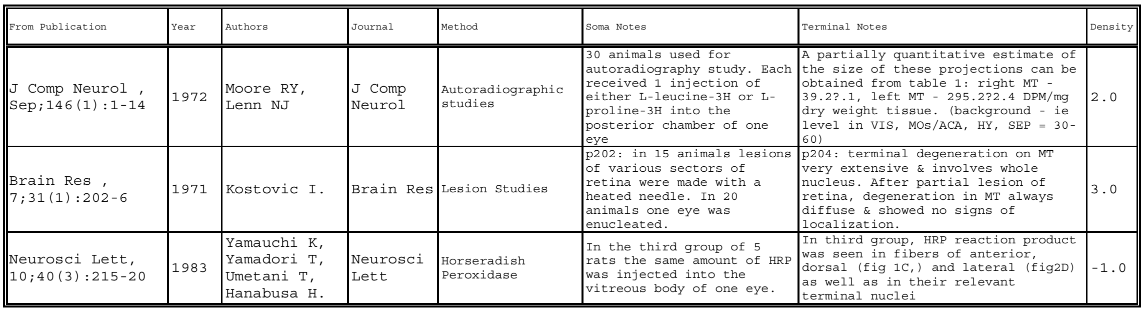
We conducted several empirical experiments within the context of two real-world application domains. Our experimental results indicate that the behaviors of the various decision fusion methods stay consistent across the two different application domains.

1. INTRODUCTION

With the advent of the World Wide Web (WWW), masses of data have suddenly become available to a large population. As a result, certain objects of common interest such as movies, research papers, or music CDs become available and are accessed frequently on the web. Meanwhile, certain people (i.e., $critics$) have the habit of (and the credential for) evaluating and reviewing these common objects and then release their evaluations. Due to the large number of common objects, people usually rely on these reviews in order to confine their access to their favorite objects. However, due to the subjectivity of the humans' perception, the set of favorite objects varies from one person to the other. That is, different persons have different degrees of confidence to various critics as well as to the various reviewing criteria. Consequently, a personalized ranking system that can order the objects by fusing the various reviews based on user's preference towards both the reviewers and the reviewing criteria is essential.

To illustrate, suppose a person wants to rank the current off-Broadway shows based on the ratings of the various components of the production. Note that these ratings are common and publicly available on some web-sites such as OOBR (http://www.oobr.com/). The reviewers assign scores to the production components (e.g., writing, directing, acting, costumes, and sound). The final ranking should combine the user's reliance on the reviewers as well as his/her preference of the reviewing criteria (in this case, the production components). A more familiar example for the research community is SIGMOD Digital Review, where research papers are reviewed by almost anyone and the typical reviewing criteria are *originality, technical correctness, presentation,* etc. In this case, a user may be interested in reading the best three papers in the area of ”decision fusion”. If the user has high confidence to certain reviewers as well as to some specific criteria (e.g., originality), then the top three papers should include those with high scores from the trusted reviewers to the $originality$ criterion.

[*] This research has been funded in part by NSF grants EEC-9529152 (IMSC ERC) and IIS-0082826, NIH-NLM R01-LM07061, DARPA and USAF under agreement nr. F30602-99-1-0524, and unrestricted cash gifts from NCR, Microsoft, and Okawa Foundation.



From Publication	Year	Authors	Journal	Method	Soma Notes	Terminal Notes	Density
J Comp Neurol , Sep;146(1):1-14	1972	Moore RY, Lenn NJ	J Comp Neurol	Autoradiographic studies	30 animals used for autoradiography study. Each received 1 injection of either L-leucine-3H or L-proline-3H into the posterior chamber of one eye	A partially quantitative estimate of the size of these projections can be obtained from table 1: right MT - 39.2?.1, left MT - 295.2?2.4 DPM/mg dry weight tissue. (background - ie level in VIS, MOs/ACA, HY, SEP = 30-60)	2.0
Brain Res , 7;31(1):202-6	1971	Kostovic I.	Brain Res	Lesion Studies	p202: in 15 animals lesions of various sectors of retina were made with a heated needle. In 20 animals one eye was enucleated.	p204: terminal degeneration on MT very extensive & involves whole nucleus. After partial lesion of retina, degeneration in MT always diffuse & showed no signs of localization.	3.0
Neurosci Lett, 10;40(3):215-20	1983	Yamauchi K, Yamadori T, Umetani T, Hanabusa H.	Neurosci Lett	Horseradish Peroxidase	In the third group of 5 rats the same amount of HRP was injected into the vitreous body of one eye.	In third group, HRP reaction product was seen in fibers of anterior, dorsal (fig 1C,) and lateral (fig2D) as well as in their relevant terminal nuclei	-1.0

Figure 1. Examples of connection reports

In this paper, we focus our attention to two very different application domains - "web metasearch fusion" and "neuroscience connection report fusion". Let us consider each application in turn.

It has been reported [22] that the search engine coverage decreases steadily as the estimated web size increases. In 1999, no search engine can index more than 16% of the total web pages. Consequently, searching data by employing only a single search engine could result in a very low retrieval rate. To solve this problem, metasearch systems, such as MetaCrawler[1], Dogpile[2], and McFind[3], are proposed to increase the search coverage by combining several search engines. Ideally, by merging various ranked results from multiple search engines into one final ranked list, metasearch systems could improve the retrieval rate. In order to improve the fusion results from metasearch systems, user preference should also be considered during the fusion process. Currently, the proposed metasearch systems only allow users to specify preference on search engines. However, the evaluation result of each page from each search engine could potentially result into several scores, each of which associates with a criterion. For example, the scores derived from the page title, URL, and the query summary would be different.

In neuroscience, neuroscientists construct theories based on many types of data. Among these types of data, the neuroanatomical connection reports may be obtained and analyzed relatively easily. Therefore, for the simplicity of the discussion, we only focus on the techniques of analyzing the neural projections. The number of neural projections between brain regions in a rat brain is in the order of $25,000$ [3]. Each projection is measured by numerous experimental specialists with different experimental methods to measure its corresponding strength. Each experimental result is documented in a *connection report*, which is associated with various criteria, such as experimenters who performed the study, the experimental methodology, the year when the experiment was conducted, and the journal where the report is published. Figure 1 illustrates the summaries of three different connection reports for the circuitry from "Retina" to "the Medial Terminal of the Accessory Optic Tract" by using three different methods. Apparently, there is a demand for computer tools that can summarize or prioritize these connection reports with the purpose of alleviating the information overload problem.

With these different application domains, we illustrate that it is necessary to incorporate user preference into the fusion process. Moreover, since each decision provided by a critic (e.g., search engine) is composed of several subdecisions (one per criterion), the fusion process requires to consider both aspects of the user preference. Therefore, the major challenge is to find an appropriate decision fusion method that can both accurately capture the users psychological behavior and efficiently aggregate the evaluations based on user preferences to the critics and the reviewing criteria.

To address both these challenges, we introduce a new concept, *two-phase decision fusion*, which aggregates the criteria evaluations in one phase and the overall weighted evaluation of the critics in the second phase. We propose several fusion operations and methods based on the previous work in information retrieval, decision making, and psychology. We also propose a learning mechanism that can minimize the user's overhead to personalize the search.

In order to evaluate the interplay of our alternative fusion operators within a fusion process, we conducted several empirical experiments focusing on our two applications of metasearch systems and the neuroscience knowledge management system. In the metasearch experiments, we utilized the data and benchmark provided by the TREC organization. In the neuroscience experiments, we used real data collected from neuroscientists. Our experimental

[1] http://www.metacrawler.com/
[2] http://www.dogpile.com/
[3] http://www.mcfind.com/

results indicate that as compared to a traditional decision fusion approach, the retrieval accuracy of our two-phase decision fusion approach is significantly increased. The most encouraging observation from our experiments is that the behaviors of the various decision fusion methods stay consistent across the two different application domains.

The remainder of this paper is organized as follows. In Section 2, we briefly describe the related work in web metasearch, neuroscience, and information retrieval. Section 3 explains the detailed design of the two-phase decision fusion process. In Section 4, we report the results of our experiments and discuss the details of the experimental setup. Section 5 concludes the paper and describes the future work.

2. RELATED WORK

2.1. Related Metasearch Work

Researchers have proposed various techniques for the metasearch systems. To date, the metasearch techniques proposed by Fox and Shaw [12] are the most adapted and effective techniques. We summarize these techniques in Table 1. Various experiments [23, 17, 12] validated that CombMNZ is the one of the most effective fusion technique. Note that most Comb{SUM,MNZ,K} methods could be consider as special cases of the aggregation operators we proposed in Section 3.2. For example, by adjusting the weights of “Weighted Voting”, the fusion process can generate the identical results as those of CombMNZ.

Technique Name	fusion result
CombMAX	maximum of input scores
CombMIN	minimum of input scores
CombMED	median of input scores
CombSUM	sum of input scores
CombANZ	average of input scores
CombMNZ	CombSUM * the number of systems retrieving this document

Table 1. Fusion techniques by Fox and Shaw

However, since metasearch systems expand the search coverage , the information overload problem could possibly be intensified. In order to improve the accuracy of returned results, researchers proposed different techniques for incorporating user preferences into metasearch systems.

The first type of personalized metasearch systems [24, 29, 8] adopt the query refinement approach. Typically, these metasearch systems modify the input query based on the corresponding user profile. Some systems[24, 8] can further select the outsourcing search engines based on user’s intent. The second types of personalized metasearch systems [28, 9] emphasize on the merging procedures. By considering user preferences during the merging process, the systems could retrieve different documents even with the same set of input lists from search engines. For example, in Inquirus 2 [9], users can assign (explicitly or implicitly) weights to different search engines and categories. The final rankings of results in Inquirus 2 are aggregated with a weighted average process. For another instance, the personalized metasearch engine proposed by Zhu et al. [28] merges the lists based on explicit relevance feedback. In this system, users can assign “good” or “bad” scores to returned pages. With content-based similarity measure, the system could evaluate final scores to all pages. Note that the importance degrees of search engines are not considered in this merging technique.

In general, most metasearch systems emphasize on one-phase merging process, i.e., the system only considers the final score of each page returned from a search engine. However, the final score provided by each search engine is composed of several similarity values, where each value corresponds to a criterion. For instance, the similarity values can be derived based on the corresponding titles of the pages, the URLs of the pages, or the summaries generated by the search engine. For another example, assume the query submitted by the user is “SARS WHO”, the metasearch system can obtain different scores from the same search engine with similar queries (e.g., “SARS WHO”, “SARS and WHO organization”, “SARS on Who magazine”, and “Severe Acute Respiratory Syndrome and WHO organization”) that are generated by a query modification process. Therefore, merging these query scores based on user preferences should also be considered.

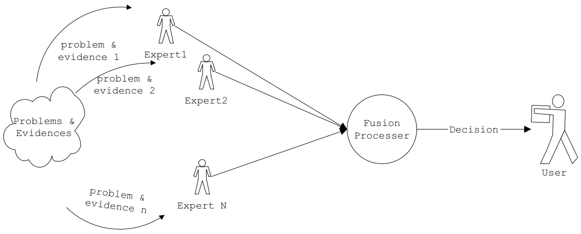


Figure 2. Processes of decision fusion

2.2. Related Neuroscience Work

Currently, we are aware of only three techniques for summarizing the connection reports in the neuroscience domain. The first one is objective relational transformation (ORT) [32], which is a relational algebra-based aggregation method. By employing a set of transformation rules, ORT can translate the connection reports across brain regions and then aggregate the transformed data of different circuitry connections. Hence, with ORT, users can retrieve the connection strength information from any user-chosen brain region. The other is the connectivity cluster analysis [4]. After transforming the connection reports to a metric space via a non-metric multidimensional scaling (NMDS) method, this technique can identify clusters of brain regions that are strongly connected with each other.

The major assumption of both these techniques is that each connection report has the same scientific quality; conversely, in practice, the qualities of connection reports vary. For example, the interpretations of experimental results by skilled experimenters are usually more accurate than those by inexperienced experimenters. In order to address this issue, the "Collation of Connectivity data on the Macaque brain (CoCoMac)" [31] system employs "precision of description codes" (PDCs) to measure the qualities of connection reports, e.g., if the figures in the reports contradict text, this report will be assigned "B" as its PDC. In this design, each connection report is associated with one and only one PDC, which describes the interpretation quality of the report. Therefore, if users have different perceptions on the same set of connection reports, they are required to assign their own PDCs to replace the old ones. Under this circumstance, obtaining customized results based on users' perceptions is time consuming for users in the CoCoMac system.

Different from ORT and NMDS that only summarize the connection reports of a given species, the connection reports employed by the NeuroHomology database [1] are across species. The NeuroHomology database allows users to retrieve the degree of homology between brain structures from two different species. In addition, to understand the brain structures of different species, users can retrieve the overall confidence level of a specific circuitry connection, which is aggregated from the corresponding connection reports with a weighted average method. In the aggregation processes of NeuroHomology, the quality of each connection report corresponds to the employed experimental technique and is assigned by the system administrator. Since the number of possible experimental techniques is limited, the overhead of obtaining customized result based on users' perceptions is less than that of CoCoMac.

2.3. Related Information Retrieval Work

The studies of decision combination, which is generally known as decision fusion, from diverse critics (from now on, we use the term "*experts*" throughout the entire paper) can be found in many research areas. For example, researchers in the forecasting community [14, 36, 11] perform decision fusion for combining different predications from experts in order to provide better forecasts; researchers in the pattern recognition community [10, 34, 35] aggregate classification results from different classifiers with the aim of improving the accuracy of recognition; researchers in the robotics community [26, 21, 18] integrate the data from different sensors with the purpose of identifying the better path for mobile robots; researchers in the web mining community [13, 33] merge the ranking results from different search engines in order to generate better query results for users.

Figure 2 provides a simple illustration of the decision-fusion processes. Basically, the system presents a problem to different experts. Each expert individually makes a decision based upon the collected evidences and provides the decision data to the central system. After receiving all the decisions, the *fusion processor* will employ the aggregation method to combine these decisions and then offer an ensemble decision to users. Note that the communication process between the central system and experts can be on-line or off-line; moreover, the decisions provided by experts can be complex objects or simple scores. Numerous studies [7, 2, 27] have demonstrated that such ensemble results are generally more accurate than the decision provided by any individual expert.

An enormous variety of aggregation methods have been presented in the literature to combine the decisions based on the problem characteristics. Some aggregation methods, such as Bagging [2], are expected to be effective when the individual decisions differ significantly from one another. Some aggregation methods, which are usually employed for the robotic and pattern recognition systems such as ENCORE [10], need to directly interact with experts online. Some aggregation methods, which are similar to the Bayesian network [26, 6], require additional data (e.g., historical information) in the aggregation process. Since the purpose of this paper is combing experts' decisions based on the preferences of human neuroscientists, the designed aggregation methods are based on psychological models.

Various studies [7, 27] observed that simple aggregation methods, such as median, average, or voting, will usually result in an effective ensemble. Some studies [25, 36, 11] even observed that average or weighted average methods could generate best decisions. Comparing to these simple aggregation methods, the empirical studies also showed that the performances of the complicated methods, such as Bayesian techniques, generally are worse than the simple methods (e.g., average or voting methods). Moreover, when the confidences of users to the ensemble decisions are concerned, psychologists [27] observed the correlation between different decisions can also affect the confidence degrees. In some cases, users are more confident when the decisions provided by experts are duplicated. In some other cases, users are more confident when the correlation between different decisions is low.

In general, most studies emphasize on one-phase decision fusion[4], i.e., each expert will directly provide one decision for the problem and the system only requires performing one aggregation to combine all decisions from experts. However, for certain applications, the decision provided by each expert is composed of several sub-decisions, where each sub-decision corresponds to a criterion (from now on we use the term "*attribute*" throughout the paper). For example, when the ice-skating referees want to evaluate the performances of skaters, they will evaluate the performances based on different attributes, such as the artistic presentation and technical skills. As another example, when a user asks for the quality information of a certain connection report from experts, it is very possible that experts have never read this report and thus can only provide some opinions on the related attributes of this connection report, such as the experimental technique employed in the connection report, the credentials of the workers who performed the experiments and interpreted the experimental results, and the journal in which the connection report appeared.

Therefore, two-phase decision fusion is required. Basically, two-phase decision fusion involves two aggregation processes, where one aggregation process combines the sub-decisions of various affecting attributes and another aggregation process ensembles the decisions of different experts. Note that each aggregation process of two-phase decision fusion should incorporate user perceptions. In other words, the final judgments will be affected not only by the confidence values of users to experts but also by the importance weights of users to attributes. To the best of our knowledge, the studies that requires two aggregation processes (such as the study by Tsikrika and Lalmas [33]) usually convert the problem from two-phase decision fusion to one-phase decision fusion. In these studies, the first aggregation process is performed offline without taking user preference into consideration. Users can only incorporate their perceptions about experts into the final aggregation process. As a result, the improvement resulted from considering more attributes as compared to that of one attribute is not significant.

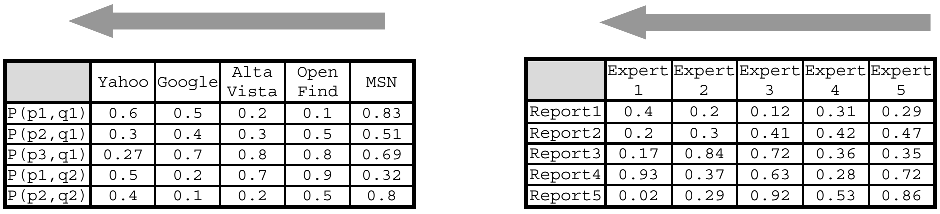
3. TWO-PHASE DECISION FUSION

In order to perform two-phase decision fusion based on user preference, the system needs to go through two main steps: aggregating the decisions based on user profiles and adjusting user profiles based on relevance feedback. We first describe the data model of two-phase decision fusion in Section 3.1. Subsequently, the aggregation techniques are described in Section 3.2. Finally, the learning method that can adjust user profiles based upon relevance feedback is described in Section 3.3.

3.1. Data Model of Two-Phase Decision

The data of traditional one-phase decision fusion can be described as a matrix. Each row represents a problem and each column represents an expert. In order to obtain the combined decision of a particular problem, the fusion processor needs to aggregate the decisions in a corresponding row. To illustrate, consider the following example.

[4] Please refer to Section 3.1 for more details.



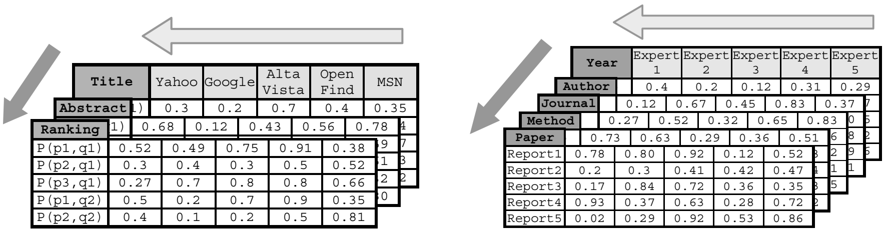
	Yahoo	Google	Alta Vista	Open Find	MSN
P(p1,q1)	0.6	0.5	0.2	0.1	0.83
P(p2,q1)	0.3	0.4	0.3	0.5	0.51
P(p3,q1)	0.27	0.7	0.8	0.8	0.69
P(p1,q2)	0.5	0.2	0.7	0.9	0.32
P(p2,q2)	0.4	0.1	0.2	0.5	0.8

a. Traditional web metasearch fusion

	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5
Report1	0.4	0.2	0.12	0.31	0.29
Report2	0.2	0.3	0.41	0.42	0.47
Report3	0.17	0.84	0.72	0.36	0.35
Report4	0.93	0.37	0.63	0.28	0.72
Report5	0.02	0.29	0.92	0.53	0.86

b. Traditional neuroscience connection report fusion

Figure 3. Example data of one-phase decision fusion



Ranking	Yahoo	Google	Alta Vista	Open Find	MSN
P(p1,q1)	0.52	0.49	0.75	0.91	0.38
P(p2,q1)	0.3	0.4	0.3	0.5	0.52
P(p3,q1)	0.27	0.7	0.8	0.8	0.66
P(p1,q2)	0.5	0.2	0.7	0.9	0.35
P(p2,q2)	0.4	0.1	0.2	0.5	0.81

a. Web metasearch fusion

Paper	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5
Report1	0.78	0.80	0.92	0.12	0.52
Report2	0.2	0.3	0.41	0.42	0.47
Report3	0.17	0.84	0.72	0.36	0.35
Report4	0.93	0.37	0.63	0.28	0.72
Report5	0.02	0.29	0.92	0.53	0.86

b. Neuroscience connection report fusion

Figure 4. Example data of two-phase decision fusion

Example 3.1: Figure 3 illustrates the data of two one-phase decision fusion applications. The data of the traditional web metasearch fusion are summarized in Figure 3.a, where $P(x, y)$ is defined as the similarity of page x to query y. In order to obtain the similarity value of page $p1$ to query $q1$, the fusion processor needs to obtain the similarity values provided from different search engines, such as Google, Yahoo, and MSN. The final similarity value of page $p1$ to query $q1$ is derived from the aggregation of the values in the first row.

Take the neuroscience connection-report aggregation as another example. Figure 3.b describes the traditional data employed by the CoCoMAC system for aggregating the qualities of connection reports, where the content of each cell represents a PDC of a connection report evaluated by an expert. The final PDC of each report is derived from the aggregation of the values in its corresponding row. ▮

On the other hand, in two-phase decision fusion problem, the decision provided by each expert is compiled from several sub-decisions where each sub-decision corresponds to an affecting attribute. Therefore, we model the data in two-phase decision fusion as *decision matrices*, where each matrix represents the sub-decisions collected from experts based on the corresponding attribute. In each matrix, the content of each cell represents a sub-decision for a problem provided by an expert based on the corresponding attribute. In order to obtain the ensemble decision of a particular problem, the fusion processor needs to perform two aggregation methods, where one aggregates the sub-decisions in a corresponding row and another one aggregates the sub-decisions in the respective matrices. The formal definition of decision matrices is given in Definition 3.1, and the illustrating example is provided in Example 3.2.

Definition 3.1: Let S denote a set of decisions. Let $\mathcal{D}_a$ be a relation from expert set $\mathcal{E}$ to the problem set Ψ for affecting attribute a, where $\mathcal{D}_a : \mathcal{E} \times \Psi \longrightarrow S$. Note that since we focus our application domain to neuroscience connection-report aggregation, the final decisions and all the sub-decisions are qualitative opinions, which can be convert to a value between $[0, 1]$ (where 0 represents that the quality of report p based on attribute a is extremely bad, and 1 represents the quality of report p based on attribute a is extremely well). Therefore, $\mathcal{D}_a(e, p) \in [0, 1]$. Choose orderings of $\mathcal{E}$ and Ψ and all decision matrices will follow the same orderings. A 2-dimensional matrix $\mathcal{M}_a$ can be used to represent this mapping.▮

Example 3.2: Consider the same applications in Example 3.1. In the web metasearch fusion problem, the similarity value of a page to a given query can be estimated based on different attributes by different search engines. Figure 4.a illustrates that the similarity values can be derived based on the ranking orders provided by the search engines, the corresponding titles of the pages, or the summaries generated by the search engines. Each matrix represents the similarity values given a specific attribute. In order to obtain the similarity value of page $p1$ to query $q1$, the fusion processor requires to perform two aggregation processes, where one process aggregates the similarity values in a corresponding row and another process aggregates the similarity values from different search engines.

Likewise, in the neuroscience connection-report aggregation problem, experts can judge the quality of each connection report based on different attributes. Figure 4.b illustrates that experts can evaluate the quality of each connection report based on different aspects, such as the publication where the report appeared, the experimental methodology employed in the report, the journal in which the report was published, the authors who wrote the report, and the year when the report was published. Each matrix represents the quality values that are judged based on a particular attribute. In order to obtain the quality of a connection report, the fusion processor needs to aggregate the quality values in the correspoding row and aggregate the values provided by different experts. ■

After acquiring the decision data from different experts, users can obtain the customized decisions, in which each decision is a weighted aggregation of the sub-decisions. Since the ensemble decisions are user dependent, we capture user preference into user profiles. A user profile is composed of three parts: *user confidence data, preferred fusion method,* and *user fuzzy cut value*. A user fuzzy cut value represents a perceptual threshold, which indicates the minimum quality value of the connection report and is employed to convert the fuzzy terms to numerical value during the aggregation process if necessary. The choice of fusion method will be described later in Section 3.2. User confidence data are employed as weights in the aggregation process, and we provide the formal definition of user confidence data as follows:

Definition 3.2: Let F denote a set of fuzzy terms. Let U represent the set of users who assign reference confidence values to the experts and assign the importance weights to the affecting attributes. π is a confidence value for a user u to an expert e; $\pi\ :\ u \in U,\ e \in \mathcal{E}\ \ \rightarrow F$. w is a confidence value (or an importance weight) for a user u to an affecting attribute a; $w\ :\ u \in U,\ a \in \Psi$. Note that the values of $\pi(u, e)$ and $w(u, a)$ are forms of human judgments and are represented as fuzzy terms.■

In practice, asking people to describe their perceptions with precise values is unreasonable. Moreover, different people have different interpretations of words. That is, the information describing personalities, preference and personal evaluation is imprecise. In order to handle this uncertainty during the query process, fuzzy logic (FL) [37] is adopted by our system. The concept of FL was first introduced by Zadeh [37] to problems for which precise formulation is not possible. The original FL has the weakness that uncertainty cannot be considered during the computation. Therefore, Karnik and Mendel advocated type 2 FL [19, 20] for overcoming this disadvantage. However, for the sake of simplicity, we only consider the original FL in this paper.

With the help of FL, our system can store and employ human's fuzzy perceptions. First, users pick up the words already defined in the system (hereafter denoted as fuzzy sets) to express their opinions. Then, all fuzzy words will be converted to real values customized for the user's perceptions according to the user's fuzzy cut value.

3.2. Fusion Methods

As described earlier, two-phase decision fusion consists of two aggregation processes, where each aggregation process involves an aggregation operator. Although there is a wide choice of aggregation operators, we only consider three operators for the two-phase decision fusion:

1. *Weighted Average:* In many empirical studies [25, 36, 11], the weighted average method usually generates the best decisions for the decision fusion problem. As a result, this method is considered as the best aggregation method in most decision fusion systems, and hence, is incorporated into our design. The formal definition of weighted average is stated below.

   Definition 3.3: Let $\mathcal{X}$ represent a set of possible sub-decisions for the problems. Let X denote a list of sub-decisions, where each element $x \in \mathcal{X}$. Let Y denote a list of importance weights, where each y_i of Y ($\sum_Y y_i$

=1) is exclusively associated with a x_i of X. The final quality value $OP_a(X, Y)$ is computed as:

$$OP_a(X, Y) = \sum_{X} y_i \times x_i \tag{1}$$

∎

2. *Weighted Voting:* From a psychological point of view, the more the confidences of the users to the final decisions, the higher the possibility that the decisions are close to the users' expectations. Psychologists [27] observed that in some cases, users are more confident when the decisions provided by experts are redundant. As a result, the voting operator can be considered as the best aggregation method in these cases, since the voting method can locate the decision that has the highest repetition rate. The formal definition of weighted voting is stated below.

   Definition 3.4: The final quality value $OP_v(X, Y)$ is computed as:

   $$OP_v(X, Y) = \arg\max_{x' \in \mathcal{X}} \{ \sum_{x_i \text{ in } X \wedge x_i = x'} y_i \} \tag{2}$$

   ∎

3. *Weighted Maximum:* Psychologists also observed that in some cases, users are more confident when the correlation between different decisions is low. Under these circumstances, considering each decision as an independent object and avoiding the merge of different decisions might achieve the highest confidence. Therefore, the maximum operator might be the most appropriate aggregation method in these cases, since each decision is handled independently in the maximum operator. The formal definition of weighted maximum is stated below.

   Definition 3.5: The final quality value $OP_m(X, Y)$ is computed as:

   $$OP_m(X, Y) = \max_{X} \{ x_i \times y_i \} \tag{3}$$

   ∎

The final decision $\lambda_{u,p}$ of problem p for user u is aggregated from decision matrices based on user profiles, which consists of the confidence values to experts, the importance weights to attributes, and the preferred fusion method. The formal fusion methods of our proposed two-phase decision fusion are stated below.

Definition 3.6: Let $X^{\mathcal{A}}_{e,p}$ denote a list of sub-decisions provided by expert e for problem p, in which each sub-decision of $X^{\mathcal{A}}_{e,p}$ can be defined as $x_a = \mathcal{M}_a(e, p), a \in \mathcal{A}$. Let $X^{\mathcal{E}}_{a,p}$ denote a list of sub-decisions corresponding to attribute a for problem p, where each sub-decision of $X^{\mathcal{E}}_{a,p}$ can be defined as $x_e = \mathcal{M}_a(e, p), e \in \mathcal{E}$. Let $Y^{\mathcal{A}}_u$ represent a list of importance weights of user u to the attributes, in which each weight of $Y^{\mathcal{A}}_u$ can be defined as $y_a = w(u, a), a \in \mathcal{A}$. Let $Y^{\mathcal{E}}_u$ represent a list of confidence values of user u to the experts, where each confidence value of $Y^{\mathcal{E}}_u$ can be defined as $y_e = \pi(u, e), e \in \mathcal{E}$.∎

Definition 3.7: The final decision $\lambda_{u,p}$ can be estimated by either Equation 4 or Equation 5. Basically, these two equations perform a very similar process. The only difference between them is the ordering of the aggregation processes, where Equation 4 aggregates the sub-decisions corresponding to different attributes first (i.e., aggregating the sub-decision in the respective matrices first) and Equation 5 aggregates the sub-decisions providing by different experts first (i.e., aggregating the sub-decision in the corresponding row first).

$$\begin{aligned}\lambda_{u,p}^{'}(e) &= OP'(X_{e,p}^{\mathcal{A}}, Y_u^{\mathcal{A}}) \\ \lambda_{u,p} &= OP''(\lambda_{u,p}^{'}, Y_u^{\mathcal{E}})\end{aligned} \tag{4}$$

or

$$\begin{aligned}\lambda_{u,p}^{'}(a) &= OP'(X_{a,p}^{\mathcal{E}}, Y_u^{\mathcal{E}}) \\ \lambda_{u,p} &= OP''(\lambda_{u,p}^{'}, Y_u^{\mathcal{A}})\end{aligned} \tag{5}$$

Note that the operators OP' and OP'' can be replaced by the weighted maximum operator OP_m, the weighted voting operator OP_v, or the weighted average operator OP_a. Therefore, there are 18 possible *fusion methods*, such as Voting_Average (in which the first aggregation operator is weighted voting and the second aggregation operator is weighted average) and MAX_Voting (which employs the weighted maximum operator first and utilizes the weighted voting operator next). Each user can specify his/her preferred fusion method for the fusion process. Alternatively, the fusion method of choice can be determined by a learning mechanism for each user based on his/her relevance feedback. ▮

The computation complexity of each aggregation operator is $O(K)$, where K is the number of sub-decisions involved in the fusion process. Consequently, the overall computation complexity of Equation 4 is $O(\|\mathcal{A}\| \times (\|\mathcal{E}\| + 1))$, where $\|\mathcal{A}\|$ is the number of attributes for the problem and $\|\mathcal{E}\|$ is the number of experts in the system, since each expert provides $\|\mathcal{A}\|$ sub-decisions and the first aggregation process would generate an aggregated sub-decision for each expert. Likewise, the overall computation complexity of Equation 5 is $O((\|\mathcal{A}\| + 1) \times \|\mathcal{E}\|)$. According to our previous study [5], the complexity of the weighted maximum can be reduced to a constant value, which is independent of the number of experts and the number of attributes. Therefore, if both aggregation operators in the fusion method are weighted maximum, the overall computation complexity of two-phase decision fusion can be reduced from $O(\|\mathcal{A}\| \times (\|\mathcal{E}\| + 1))$ to $O(1)$.

Moreover, if both aggregation operators in the fusion method are the weighted average operators, the fusion processor can replace two aggregation processes by one aggregation process and still obtain the identical result. This one-step aggregation method is defined as follows:

Definition 3.8: The final decision $\lambda_{u,p}$ is computed as:

$$\lambda_{u,p} = \sum_{a\in\mathcal{A}, e\in\mathcal{E}} \{w(u,a) \times \pi(u,e) \times \mathcal{M}_a(e,p)\} \tag{6}$$

▮

Similarly, if both aggregation operators in the fusion method are the weighted maximum operators, the two-phase fusion method can be replaced by a one-step aggregation method and still obtains the same result.

3.3. The Learning Method

In order to perform two-phase decision fusion based on user preference, the fusion method heavily relies on user profiles. In Section 3.2, we assumed that users would supply accurate user profiles. However, in practice, obtaining user profiles has been challenging. For example, users may be too busy to provide the data or they might unintentionally input the incorrect information. A more appropriate approach should offer a learning mechanism to correct these errors. Building on this premise, we utilize the users' relevance feedback thus generating a better user profile automatically using a genetic algorithm (GA).

GA [15] is an iterative search technique based on the spirit of natural evolution. By emulating biological selection and reproduction, GA can efficiently search through the solution space of complex problems, where each candidate solution is represented by a chromosome and evaluated by a fitness function for the survival chance. Although GA consists of many components, such as the mutation operator and the crossover operator, only the fitness function and the coding/decoding method for chromosomes are required to be specially designed per each application.

Our proposed learning mechanism, named *GADiF (GA-based learning mechanism for DecIsion Fusion)*, is an automatic and off-line process that is activated after receiving user relevance feedback. User involvement is only needed for providing the relevance feedback as the goal of GA prior to the beginning of evolution. GADiF employs GA for generating user profiles, which is composed of a preferred fusion method, user confidence data, and a user fuzzy cut value, by decoding the best chromosome to replace the existing profile after its evolution. The objective of GA is locating the best user profile that can achieve the closest fusion result to user feedback. In other words, by employing the located fusion method and the located user preference, the fusion process can generate the ensemble decision closest (or almost closest) to the user feedback.

Subsequently, we explicate the coding design of GADiF for GA. The chromosomes represent possible user profiles for a specific user, and each gene in the chromosome corresponds to a weight value or an aggregation operator. Four types of genes are involved in the chromosomes (see Figure 5). One is user confidence information with $\|\mathcal{E}\|$ genes, where $\|\mathcal{E}\|$ is the number of experts in the system. The value of the ith gene is an integer in $[0, L-1]$, where L is the number of fuzzy terms used in the system, and denotes the user's confidence level to expert i. The second one is the importance weighted information with $\|\mathcal{A}\|$ genes, where $\|\mathcal{A}\|$ is the number of attributes associated with the fusion process.

The third type of genes corresponds to the fusion method. The value j of the $\|\mathcal{E}\| + \|\mathcal{A}\| + 1$th gene represents the selection of the aggregation operator OP', where $(jMOD3) = 0$ denotes the operator is the weighted average operator, $(jMOD3) = 1$ denotes the operator is the weighted voting operator, and $(jMOD3) = 2$ denotes the operator is the weighted maximum operator. The value of the $\|\mathcal{E}\| + \|\mathcal{A}\| + 2$th gene represents the selection of the aggregation operator OP''. The value $j3$ of $\|\mathcal{E}\| + \|\mathcal{A}\| + 3$th gene represents the preferred fusion function, where $(j3MOD2) = 0$ denotes Equation 4 is the preferred function, and $(j3MOD2) = 1$ denotes Equation 5 is the preferred function. The fourth type of the gene is a user fuzzy cut value, whose value is $(t+1)/L$, where $t \in [0, L-1]$ is the value of this gene. Figure 5 illustrates the coding design. By employing this coding design, GADiF can locate the best (or the nearly best) combination of user preference and the preferred fusion method after evolution.

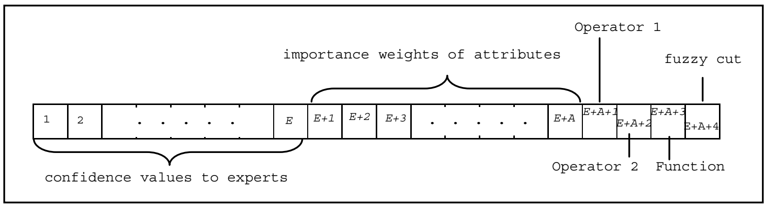


Figure 5. The coding design of chromosomes

Next, we describe the fitness function of GADiF, which heavily utilizes the users' relevance feedback. Users can provide their expected decision values by adjusting the output values. The fitness function first decodes the chromosome into a user profile. Then, it performs a two-phase decision fusion based on this profile using Definition 3.7. In other words, this process needs to interact with the system for obtaining the ensemble decisions. Finally, it generates the fitness value by measuring the similarity between the ensemble decision and the users' relevance feedback. The fitness functions could be varied depended on the application domains.

In the neuroscience connection report fusion, the fitness value could be based on the average absolute error. The average absolute error (AAE) is computed by Equation (7)

Definition 3.9: Let H_o be the original score list by a user and let H_q be the decision list for a set of problems generated by aggregation methods. The average absolute error is:

$$\text{AAE}(H_o, H_q) = \frac{\sum_i^N (H_o(i) - H_q(i))}{N} \tag{7}$$

∎

In the web metasearch systems, the fitness values can be evaluated based on the relevance feedback or estimated from navigation behaviors without explicit acquisition of user relevance feedback, given the assumption that users only navigate potentially desired items and they only browse the pages of uninteresting items for a comparably shorter time. The detailed description could be found in our previous work [30].

Once a user offers his/her user feedback to trigger the learning process, GADiF first encodes the corresponding user profile to a chromosome and randomly generates other chromosomes as the initial population. Subsequently, GA iteratively discovers better user profiles until it achieves the terminal condition such as the fitness value of one chromosome being 0 or the generation number being 50. Finally, GADiF decodes the best chromosome to a user profile and replaces the current user profile for the two-phase fusion processor.

4. PERFORMANCE EVALUATION

We conducted numerous experiments for comparing the different fusion methods in two different application domains. The experiments focused on the web metasearch fusion are described in Section 4.1, and the experiments emphasized on the neuroscience connection report fusion are discussed in Section 4.2. The experiments are implemented in C language[5] and run on a Pentium III 600 MHZ processor with Microsoft Windows 2000.

4.1. Web Metasearch

4.1.1. Experimental Methodology

In this experiment, we employ the systems submitted to TREC-2001 Web Track for the topic relevance task as our search engines. These search engines work on a 10 gigabyte, 1.69 million page collection. The topic relevance task employs queries taken from real web search logs. There are 50 topics (topic 501-550) for this task. Each topic has its correct answers, which are documents received binary judgments ("relevant" or "irrelevant"). Note that each document receives one-and-only-one judgment regardless the user submitted the query. Each search engine returns 1000 documents and their corresponding scores for each topic. Notice that these search engines are compared only on the basis of title-only queries. In addition to the title data, search engines are allowed to utilize three types of data: links, document structure, and URL text.

Among 28 official submission groups, we randomly select 10 groups as outsourcing search engines, where the search engine would have higher selectivity if it generates several submissions based on different types of data. For example, Yonsei would have a higher selectivity than IBM-Haifa, since Yonset provided two different submissions (where one employed document structures and links, and the other only utilized document structures) while IBM-Haifa provided four similar submissions (where each of them use both document structures and URL text). If the selected group has several submissions where each submission employs the same types of data, only one submission will be randomly selected. Table 2 depicts the accuracy of these selected search engines[6].

	Precision					Recall				
	number of documents retrieved					number of documents retrieved				
	5	10	20	50	100	5	10	20	50	100
Average	0.169	0.147	0.127	0.113	0.105	0.017	0.030	0.052	0.128	0.215
Max	0.256	0.19	0.155	0.133	0.17	0.030	0.046	0.066	0.156	0.26
Min	0.1	0.096	0.092	0.086	0.080	0.007	0.014	0.028	0.081	0.141
STD	0.039	0.030	0.021	0.017	0.016	0.007	0.010	0.012	0.025	0.035

Table 2. Precision and Recall of original search engines

We consider the types of data employed in the search engines as our "affecting attributes". As a result, the score of a document in a submission will be treated as the scores based on one attribute. After selecting all search engines, GADif (see Section 3.3) is triggered for locating a possible user profile that can achieve the best result (i.e., the precision is the highest during the entire search process) for each query. During the fusion process, the system selects the top 100 documents from each submission and generates the top 20 documents for each query.

[5] We developed GADiF using SUGAL [16] for its wide range of operators and data types.

[6] Note that the precision and recall data are evaluated based on the resulting files we obtained from the TREC web site.

4.1.2. Experimental Results

It should be noticed that the relevance judgments provided by TREC are not the "good" relevance feedback in this experiment for two reasons. First, they are not user dependent (i.e., the judgments for each query are not obtained from a single user but from several users). Second, they are binary judgments, where the ideal judgments should be able to distinguish how relevant the documents are to the query. As a result, the experimental results could be biased.

The results shown for each set of experiments are averaged over many runs. Each run is executed with different seeds for the random number generator functions. The coefficient of variance across these runs was smaller than 0.1%, which shows the results are independent of the specific run and are independent of any specific query. We conducted several sets of experiments to compare the different fusion methods. In these experiments, we observed that Equation 4 outperforms Equation 5 in most cases; however, the difference between the two is insignificant. Hence, to simplify the results, we only report the results for Equation 4, i.e., the sub-decisions corresponding to different attributes will be aggregated first and the ensembled decisions of experts will be aggregated subsequently.

	Two-Phase		One-Phase		Improvement of two-phase decision over	
	Precision	Recall	Precision	Recall	the best input	one-phase decision
@5	0.526	0.103	0.48	0.090	105.65%	9.68%
@10	0.490	0.169	0.43	0.152	157.68%	13.86%
@20	0.484	0.292	0.42	0.259	212.39%	15.29%

Table 3. Comparison of one-phase decision fusion and two-phase decision fusion

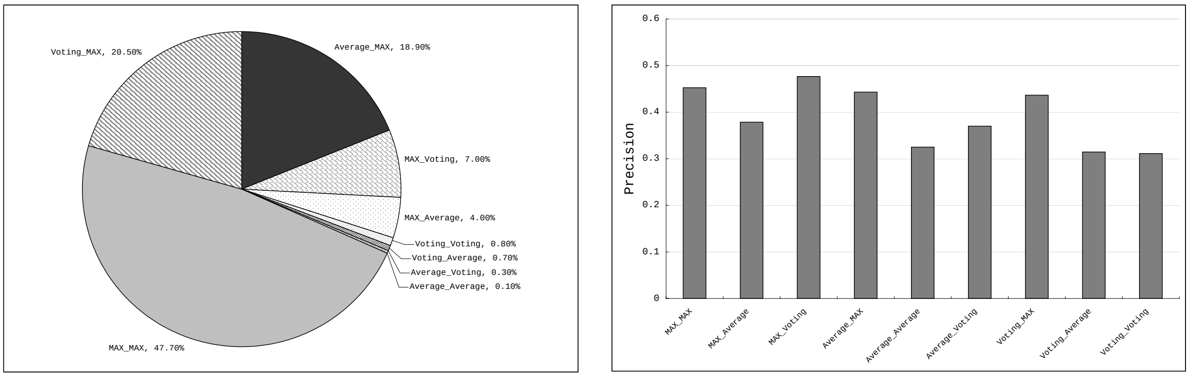
In the experiments of Table 3 and Figure 6.a, our proposed learning mechanism, GADiF, identified the fusion method that can achieve the best results based on the user preferences on search engines and attributes. Assume that the user preferences obtained by GADif are accurate for two-phase decision fusion and one phase decision fusion processes. In Table 3, the forth and fifth columns shows the performances of a traditional one-phase decision fusion process, i.e., each search engine only provides one score for each document (where the score comes from its best submission).

Table 3 illustrates that the two-phase decision fusion outperforms the traditional one-phase decision fusion even though the relevance judgments provided by TREC are not suitable for two-phase decision fusion. As compared to the best input (described in Table 2), the precision of two-phase decision fusion achieve over 200% improvement when 20 documents are retrieved. The improvement rates of two-phase decision fusion over one-phase decision fusion range from 9% to 15% depending on the number of documents retrieved.

Figure 6.a illustrates that if all user preferences are applicable and accurate, the fusion method that employs two maximum operators (termed *MAX_MAX*) can achieve the best result with a 48% probability. Under the same circumstance, the fusion method that employs two average operators (termed *Average_Average*) only can outperform other fusion methods with a 0.1% probability. Although the experimental results of Figure 6.a can illustrate the probability of achieving the best result for each fusion method, these results alone cannot validate which fusion method is superior to others. It is still possible that Average_Average is the best choice, despite the fact that Average_Average only surpasses other fusion methods with a extremely low probability in Figure 6.a. For example, if Average_Average surpasses other methods with a significant difference and is only defeated with a negligible difference, Average_Average could achieve the best overall performance and hence be the best selection. Therefore, we performed the experiment of Figure 6.b to compare the overall performance of each fusion method in order to clarify this issue.

Figure 6.b compares the overall performances of different fusion methods. The Y-axis of Figure 6.b depicts the precision of the final decisions. In the experiment of Figure 6.b, we slightly modified the design of GADiF; i.e., GADiF only locates the user preferences on search engines and attributes according to the specified fusion method, it does not, however, suggest the best aggregate operators anymore.

As revealed by Figure 8.b, the overall performance of the fusion method that employs two maximum operators first and utilizes a voting operator next (termed *MAX_Voting*) is unexpectedly superior to other fusion methods. As we mentioned earlier, the weighted voting method is a general case of CombMNZ proposed by Fox and Shaw [12]. Moreover, since the relevance judgments provided by TREC are assembled from several people (probably with the concept of one-phase decision fusion), it is not surprising that MAX_Voting outperforms other methods. Furthermore, MAX_MAX still outperforms most fusion methods, except the MAX_Voting method. This indicates the MAX_MAX method is preferred for two-phase decision fusion as compared to most alternative methods we proposed. The further discussion on the behaviors of fusion methods are provided in Section 4.2.2.



Note: X_Y notation represents the first aggregation operator is X and the second aggregation operator is Y

a. Percentages of times that each fusion method achieves the best result

b. Performance comparison between different fusion methods

Figure 6. Comparison of different fusion methods

4.2. Neuroscience Connection Report Fusion

4.2.1. Experimental Methodology

In these experiments, we employed the fusion methods described in Section 3.2 to obtain the ensemble qualities of connection reports. The expected qualities of the connection reports are obtained from neuroscientists via a website or a questionnaire. Based on our past experiences, the data provided by non-domain experts are usually inconsistent, and consequently, the bias caused by the inconsistency could lead to an incorrect conclusion. In order to prevent the incorrect conclusion, we insisted on acquiring opinions (i.e., the expected qualities of the connection reports) only from neuroscientists who specialize in neuroanatomical circuits, even though the number of domain experts who could participate can be much smaller than that of non-domain experts.

The participants for this survey were 15 neuroscientists who familiar with neuroanatomical circuits of rat brain. Three of them assigned their sub-decisions corresponding to certain attribute values on our demonstration website[7]. These three participants are considered as experts, who provide the sub-decisions of problems, in our experiments. The other 12 neuroscientists provided their judgments of different connectivity reports by answering a designed questionnaire. The answers of the questionnaire are employed as user expectations of the decision fusion results. This questionnaire consists of 44 connectivity reports associating with 9 different circuits in the rat brain. The connectivity reports are randomly selected from a medium-sized database[8] that holds roughly 6000 connection reports, which involve 7 different experimental techniques and are collected from 1200 papers. Each connection report is associated with five attributes: experimental technique, author, paper, published journal, and published year. The participants were asked to evaluate each connectivity report by assigning a score between 0 to 1, where 0 represents the report is extremely unreliable and 1 represents the report is highly credible. Note that the 12 people who answered the questionnaire only provide one and only one quality value to each connection report.

Figure 7 illustrates the benchmarking processes for the experiments in each run. Among the 12 sets of questionnaire answers, at each run, two of them are randomly selected as training data and others are used as testing data. Subsequently, we expanded the number of experts from 3 to 10 based on the real expert data and training data with a specially designed learning mechanism[9]. This learning mechanism would simulate the sub-decisions corresponding to different attributes for the simulated experts.

After having all expert data (where some of them are real data from neuroscientists and some of them are generated by the learning mechanism), GADiF is triggered for locating a possible user profile that can achieve the best result

[7] http://formosa.usc.edu/NeuroScholar/login.html

[8] The data are obtained from a experimental website related to NeuroScholar, which is a neuroscience knowledge management system (http://neuroscholar.usc.edu/).

[9] Note that this learning mechanism is different from GADiF proposed in Section 3.3. The objective of this mechanism is generating 7 simulated experts and their corresponding sub-decisions.

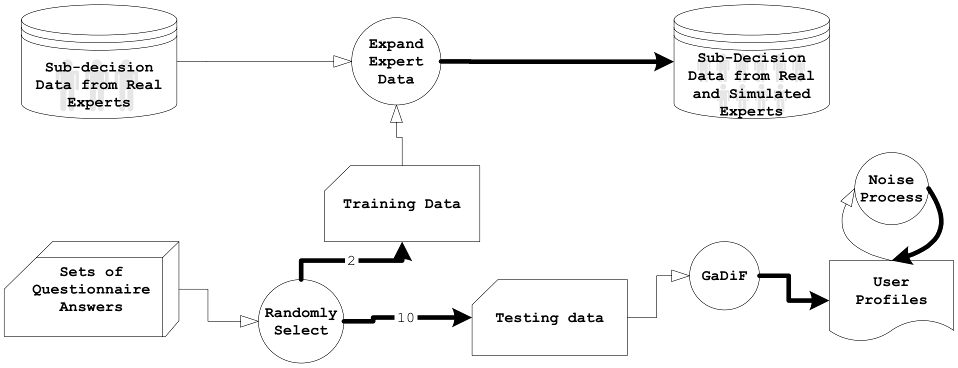


Figure 7. Benchmarking processes in each run

(i.e., the absolute error between the ensemble qualities and the expected qualities from neuroscientists is the minimum during the entire search process) for the testing data.

To simulate imperfect or incomplete user profiles, the system tunes the user profiles by a noisy process according to a noise level. The noisy process employs a random number generator, which is based on the linear congruent method, and the current system clock time as the seed. For each element in the user profiles, the noisy process adds a noise value in the interval $[-(\text{noise level})/2, (\text{noise level})/2]$. Noise level 0 represents perfect user profiles and noise level 10 represents complete chaos. The notion of *average absolute error* defined in Equation 7 is used to evaluate the accuracy.

4.2.2. Experimental Results

In this section, we emphasize on two sets of experiments and their corresponding results. In the first set, we compare the performances of different fusion methods when the user information is incomplete (i.e., the confidence values to experts or the importance weights on attributes are absent). With the second set, to determine the fusion method that can provide the best performance when the user information is complete, we examine various setups and discuss the performances.

The results shown for each set of experiments are averaged over many runs. Each run is executed with different seeds for the random number generator functions. In each run, two sets of answers to the questionnaire were randomly selected as training data for generating the simulated experts, and the other 10 sets of answers were employed as testing data for comparing the performances of fusion methods. The coefficient of variance across these runs was smaller than 2.5%, which shows the results are independent of the specific run and are independent of any specific user. Similar to Section 4.1.2, we also observed the difference between Equation 4 and Equation 5 is marginal, although Equation 4 once again outperforms Equation 5. To simplify the results, we only report the results for Equation 4.

4.2.2.1. Performance Analysis of the Incomplete-User-Information Cases

	Operator 1		
Operator 2	MAX	Average	Voting
MAX	0.3946	0.3780	0.3851
Average	0.3831	0.3712	0.4011
Voting	0.3832	0.37715	0.4153

The numerical values are AAEs

Table 4. Comparison of different fusion methods (without user preferences)

Table 4 depicts the performances of different fusion methods without user preferences, i.e., the system equally assigned a default weighting value for both the confidence values to experts and the importance weights on attributes. The performance value is measured by AAE of the ensemble decisions. According to Table 4, the fusion method that employs two weighted average operators, termed Average_Average, can achieve the best results when the user preferences are unavailable. Moreover, as revealed by Table 4, the fusion methods that comprise the average operator constantly surpass other fusion methods. Consistent with the conclusion of many other studies [25, 36, 11], the results of this experiment once again demonstrate that the simple average method can be the most appropriate fusion method even for the two-phase fusion problem when the user preferences are unknown.

In Table 5, the first row shows the performances of three aggregation operators in a traditional one-phase decision fusion problem, i.e., each expert only provides one decision for each problem. The second row illustrates the performances of three aggregation operators in the two-phase decision fusion when the user preferences on attributes are absent. All aggregation operators incorporate the confidence values of users to experts, which are provided by GADiF. The performance value is also measured by AAE of the ensemble decisions.

	Employed operator		
	MAX	Average	Voting
Traditional one-phase decision fusion	0.273347879	0.148561069	0.240375518
Two-phase decision fusion without user preferences on attributes	0.200794813	0.159147222	0.22339541

The numerical values are AAEs

Table 5. Comparison of one-phase decision fusion and two-phase decision fusion

Table 5 indicates the performance of the average operator is better than those of voting and maximum operators when experts only provide one decision for each problem or when user preferences on attributes are ignored. Table 5 also illustrates that although the two-phase decision fusion incorporates more decisions in the fusion process, the improvement resulted from considering more decisions per problem as compared to one decision per problem is marginal.

Moreover, we observe that performances of the average operator decline when the fusion process incorporates more data[10]. This may contribute to the sub-decision disagreements. When the additional attributes are added, the number of decisions corresponding to a specific problem grows and hence the possibility of sub-decision disagreements increases. Since the average operator is not designed for aggregating diverse values, the aggregated sub-decisions for the second phase aggregation process will more likely converge to a mean value when using the average operator. Unlike the average operator, the maximum and voting operators do not combine the decisions and therefore will be impacted less due to sub-decision disagreements.

4.2.2.2. Performance Analysis of the Complete-User-Information Cases

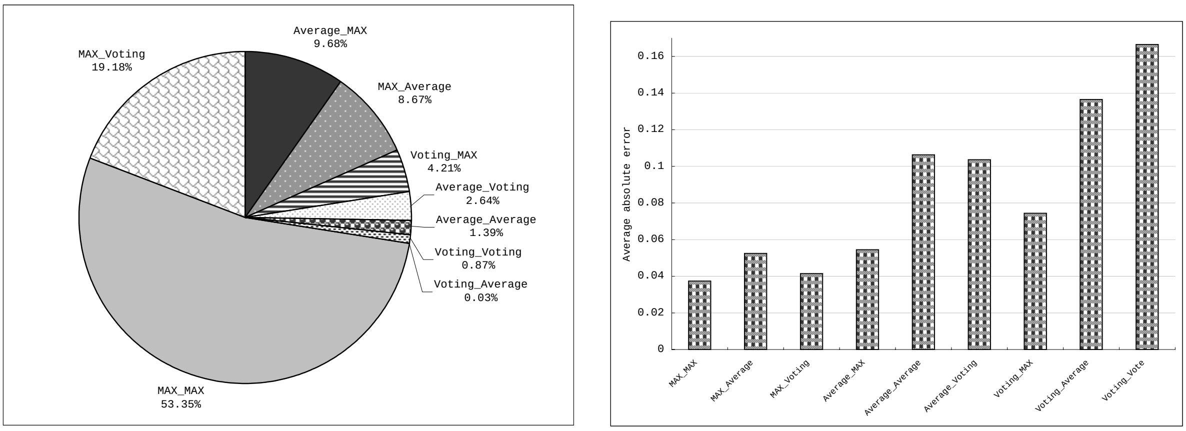
In the experiment of Figure 8.a, our proposed learning mechanism, GADiF, identified the fusion method that can achieve the best results based on the user preferences on experts and attributes. Figure 8.a compares the percentages of times that each fusion method was selected by GADiF. Note that the results are averaged over thousands of runs, where each run is executed with different seeds for the random generator function, i.e., the training and testing data sets are different in each run, thus the result is independent of the users.

Assume that the user preferences obtained by GADiF are accurate. Figure 8.a once again illustrates that if all user preferences are applicable and accurate, MAX_MAX can outperforms other fusion methods with the highest probability (53%). Similar to the observation in Section 4.1.2, Average_Average only can outperform other fusion methods with a tiny probability (1.4%) even though Average_Average usually outperforms others when the user preferences are incomplete.

The objective of the experiment in Figure 8.b is to compare the overall performances of different fusion methods. The Y-axis of Figure 8.b depicts the average absolute error of the final decisions as compared to the user opinions computed by Equation (7). Again, in the experiment, GADiF only locates the user preferences on experts and attributes according to the specified fusion method; it does not, however, suggest the best aggregate operators anymore.

As revealed by Figure 8.b, the overall performance of the MAX_MAX method is superior to other fusion methods. Except MAX_MAX, MAX_Voting outperforms other fusion methods. Surprisingly, after incorporating user

[10] Although many studies or many models suggested that additional data should improve the accuracy of results, few studies reported the similar observations. For example, Rantilla and Budescu [27] discovered that decision makers would have higher confidence if fewer experts provided the decisions; Chen and Shahabi [5] reported that in image retrieval, results would be less accurate if more physical-feature comparison functions were utilized.

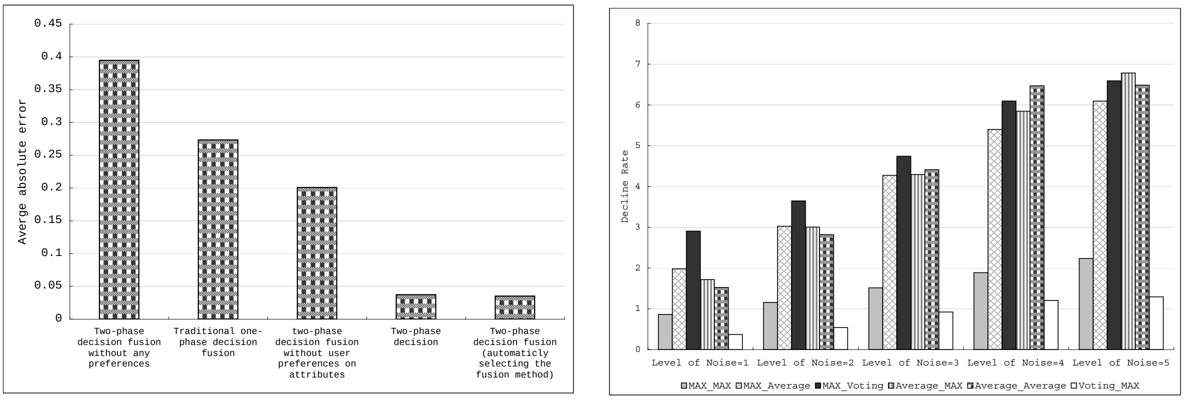


Note: X_Y notation represents the first aggregation operator is X and the second aggregation operator is Y

a. Percentages of times that each fusion method achieves the best result

b. Performance comparison between different fusion methods

Figure 8. Comparison of different fusion methods



a. Impacts of data quantity

b. Impacts of noise

Figure 9. Performance comparison of different scenarios

preferences on attributes, using the average operator usually results in less accurate decisions (e.g., the performance of Average_Average was surpassed by most fusion methods in Figure 8.b.), while the average operator outperformed other operators in the experiments of Section 4.2.1. Notice that the observation in Figure 8 is extremely similar to that in Figure 6. This indicates that the behaviors of the various decision fusion methods stay consistent across the two different application domains

Once again, the inferior performance of Average_Average may be possibly due to the impact of sub-decision disagreements. Although incorporating the user preferences on attributes can improve the decisions, the sub-decision disagreements for the average operator still hold. Accordingly, the performance of MAX_Average, which employs a maximum operator first and subsequently utilizes an average operator, is better than that of Average_Average, because utilizing the maximum operator can reduce the effect of sub-decision disagreements. From a psychological point of view, employing experts' decisions implies users trust of experts' professional knowledge. The more manipulation on the experts' opinions, the higher the distrust of users to experts. Consequently, by performing more manipulation on experts' opinions, the fusion process not only damages experts' input but also decays reliance of users on experts. Therefore, the fusion methods that can preserve experts' knowledge would be closer to user preferences in the real world.

Subsequently, to compare the impacts of the amount of data incorporated in the fusion process, Figure 9.a summarizes the results of previous experiments. The Y-axis of Figure 9.a represents the average absolute error. Except for the last bar, each bar of Figure 9.a denotes the performance of MAX_MAX method in a specified scenario. To illustrate, the first bar represents the overall performance of MAX_MAX in the two-phase decision fusion without any user preferences; the second bar denotes the overall performance of MAX_MAX in the one-phase decision fusion (i.e., each expert only provides one decision for each problem) with user preferences on experts; the third bar demonstrates the overall performance of MAX_MAX in the two-phase decision fusion when only user preferences on experts are available; the fourth bar represents the overall performance of MAX_MAX in two-phase decision fusion when the user preferences on experts and attributes are applicable. The last bar of Figure 9.a shows the overall performance of Figure 8.a, where the two-phase fusion processor utilizes all user preferences and the fusion method is selected by GADiF.

As revealed by Figure 9.a, adding user preference data on experts or attributes can achieve higher improvement than increasing the number of sub-decisions. For example, compared to the performance of only employing user preferences to experts, employing user preferences on both experts and attributes can achieve an 80% improvement, while employing additional sub-decisions can only achieve a 27% improvement (which is comparable to the performance of the traditional one-phase decision fusion). Therefore, the reason that the improvement resulted from considering more decisions as compared to one decision per problem is not significant in some studies (such as the study of web search fusion by Tsikrika and Lalmas [33]) could probably be attributed to the lack of user preferences information on attributes. Moreover, although selecting the appropriate fusion method based on the user preferences can enhance the decision accuracy, the improvement is insignificant (i.e., a 5% improvement). This result indicates that the MAX_MAX method can be considered as the default fusion method in two-phase decision fusion problem while the user preferences are applicable and accurate.

Finally, in order to compare the fusion performances when the user preference data are imperfect, we introduced five different noise levels as illustrated in Figure 9.b. Note that we only illustrate the performances of fusion methods that can achieve better accuracy, hence the performances of Voting_Voting, Voting_Average, and Average_Voting are not included in Figure 9.b. The Y-axis denotes the decline rate calculated by Equation (8), where N_o is the performance of two-phase decision fusion process while the user preferences are perfect, and N_i represents the performance when the user preferences are defective.

$$\text{Decline Rate}(N_o, N_i) = \frac{(N_o - N_i)}{N_0} \tag{8}$$

As observed in Figure 9.b, Voting_MAX is the most robust fusion method when the user preferences are imperfect. Although the performance of MAX_MAX is not as robust as that of Voting_MAX, MAX_MAX significantly outperforms the rest of the fusion methods. Moreover, as shown in Figure 9.b, the noise severely impacts the performance of Average_Average. This observation provides another reason that the Average_Average method is not the best choice in many two-phase decision fusion problems.

5. CONCLUSION AND FUTURE WORK

In this paper, we introduced a new concept, *two-phase decision fusion*, where opinions of the experts are aggregated based upon user perceptions on experts and attributes. For the two-phase decision fusion process, we proposed several aggregation methods as well as a learning mechanism.

We conducted several experiments with two real-world applications. A summary of our main observations are as follows. We concluded that if user preferences are accurately available, the MAX_MAX method is the superior fusion method for several reasons. First, it is less sensitive to sub-decision disagreements. Second, it can achieve the best results with a high probability and almost always (except for one case) has the best overall performance. Third, it is robust to noise. Finally, as discussed in Section 3.2, the worst-case computation complexity of MAX_MAX can be reduced to a constant value, which is independent of the number of experts and attributes, while the computation complexities of other fusion methods depend on the number of experts and attributes. On the other hand, if the system can only rely on the user preference on either the expert or the attribute (but not both), our fusion process is reduced to a one-phase decision fusion method. In this case the MAX_Voting method becomes the superior technique. Finally, if the user preferences are totally inaccessible, then the Average_Average method generates as good of a result as one can achieve.

As our future work, we intend to investigate an adaptive two-phase decision fusion system. In this case, the system may employ Average_Average to aggregate decisions for new users. Subsequently, after obtaining enough user feedback, the system may switch to either the MAX_MAX method or MAX_Voting.

References

[1] M. Bota and M. A. Arbib. The neurohomology database. In *Michael A. Arbib, and Jeffrey Grethe, (Editors) Computing the Brain: A Guide to Neuroinformatics*, pages 337–351. Academic Press, 2001.

[2] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[3] G. A. Burns. Knowledge mechanics and the neuroscholar project: A new approach to neuroscientific theory. In *Michael A. Arbib, and Jeffrey Grethe, (Editors) Computing the Brain: A Guide to Neuroinformatics*, pages 319–335. Academic Press, 2001.

[4] G. A. P. C. Burns and M. P. Young. Analysis of the connectional organization of neural systems associated with the hippocampus in rats. *Philosophical Transactions: Biological Sciences*, 355:55–70, 2000.

[5] Y.-S. Chen and C. Shahabi. Yoda, an adaptive soft classification model: Content-based similarity queries and beyond. *ACM/Springer Multimedia Systems Journal*, 8(6):523–535, 2003.

[6] R. Clemen and R. Winkler. Combining economic forcasts. *Journal of Business and Economic Statistics*, 4:39–46, 1986.

[7] R. T. Clemen. Combining forcasts: A review and annotated bibliography. *International Journal of Forecasting*, 5:559–583, 1989.

[8] S. L. W. B. A. K. C. G. E. Glover, G.W. Flake and D. Pennock. Improving category specific web search by learning query modifications. In *Proceedings of Symposium on Applications and the Internet*, 2001.

[9] W. B. E. Glover, S. Lawrence and C. Giles. Architecture of a metasearch engine that supports user information needs. In *Proceedings of Eighth International Conference on Information and Knowledge Management (CIKM'99)*, 1999.

[10] M. Fairhurst and A. Rahman. Enhancing consensus in multiple expert decision fusion. *IEE Proceedings on Vision, Image and Signal Processing*, 147:39–46, 2000.

[11] I. Fischer and N. Harvey. Combining forecasts: What information do judges need to outperform the simple average. *International Journal of Forecasting*, 15:227–246, 1999.

[12] E. Fox and J. Shaw. Combination of multiple searches. In *Proceedings of the Second Text REtrieval Conference (TREC-2)*, pages 243–249, 1994.

[13] V. Gorodetski, O. Karsaev, and V. Samoilov. Data fusion and semantic web: Meta-models of distributed data and decision fusion. In *Proceedings of the International Workshop Semantic Web Mining (PKDD'02)*, Helsinki, August 2002.

[14] C. W. J. Granger. Combining forcasts – twenty years later. *Journal of Forecasting*, 8:167–173, 1989.

[15] J. Holland. *Adaption in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, Michigan, 1975.

[16] A. Hunter. Sugal programming manual. *http://www.trajan-software.demon.co.uk/sugal.htm*, 1995.

[17] J.H.Lee. Combining multiple evidence from different propertices of weighting schemes. In *Proceedings of the 18th ACM SIGIR*, pages 180–188, 1995.

[18] G. Kamberova, R. Mandelbaum, and M. Mintz. Statistical decision theory for mobile robotics: theory and application. In *Proceedings of 1996 IEEE/SICE/RSJ International conference on multisensor fusion and integration for intelligent systems*, Washington DC, December 1996.

[19] N. Karnik and J. Mendel. Introduction to type-2 fuzzy logic systems. In *Proceeding of 1998 IEEE FUZZ Conference*, pages 915–920, Anchorage, AK, May 1998.

[20] N. Karnik and J. Mendel. Operations on type-2 fuzzy sets. *Int'l. J. on Fuzzy Sets and Systems*, 2000.

[21] M. A. A. L.A. Gee, C. Dumont. Multisensor fusion for decision-based control cues. In *Proceedings of SPIE Vol. 4052 Signal Processing, Sensor Fusion, and Target Recognition IX*, Prlando, Florida USA, April 2000.

[22] S. Lawrence and C. Giles. Accessibility of information on the web. *Nature*, 400:107–109, 1999.

[23] J. Lee. Analyses of multiple evidence combination. In *Proceedings of the 20th ACM SIGIR*, pages 267–275, 1997.

[24] D. Z. M. Chau and H. Chen. Personalized spiders for web search and analysis. In *Proceedings of ACM/IEEE Joint Conference on Digital Libraries*, 2001.

[25] L. Maines. An experimenntal examination of subjective forecast combination. *International Journal of Forecasting*, 12:223–233, 1996.

[26] B. Moshiri, M. R. Asharif, and R. HoseinNezhad. Pseudo information measure: a new concept for extension of bayesian fusion in robotic map building. *Information Fusion*, 3:51–68, 2002.

[27] A. K. Rantilla and D. V. Budescu. Aggregation of expert opinions. In *Proceedings of the 32nd Hawaii International Conference on System Sciences*, Hawaii, January 1999.

[28] K. C. S. Zhu, X. Deng and W. Zheng. Using online relevance feedback to build effective personalized metasearch engine. In *Proceedings of Second International Conference on Web Information Systems Engineering (WISE'01)*, 2001.

[29] A. Scime and L. Kerschberg. Websifter: An ontology-based personalizable search agent for the web. In *Proceedings of International Conference on Digital Libraries: Research and Practice*, 2000.

[30] C. Shahabi and Y.-S. Chen. An adaptive recommendation system without explicit acquisition of user relevance feedback. *To Appear in the Distributed and Parallel Databases Journal*, 2003.

[31] K. E. Stephan, L. Kamper, A. Bozkurt, G. A. P. C. Burns, M. P. Young, and R. Kotter. Advanced database methodology for the collation of connectivity data on the macaque brain (cocomac). *Philosophical Transactions: Biological Sciences*, 356:1159–1186, 2001.

[32] K. E. Stephan, K. Zilles, and R. Kotter. Coordinate-independent mapping of structural and functional data by objective relational transformation (ort). *Philosophical Transactions: Biological Sciences*, 355:37–54, 2000.

[33] T. Tsikrika and M. Lalmas. Merging techniques for performing data fusion on the web. In *Proceedings of the 2001 ACM CIKM International Conference on Information and Knowledge Management*, Atlanta, Georgia, USA, November 2001.

[34] P. Verlinde and G. Chollet. Comparing decision fusion paradigms using k-nn based classifiers. In *Proceedings of In Second International Conference on Audio- and Videobased Biometric Person Authentication (AVBPA)*, Washington D. C., USA, March 1999.

[35] N. M. Wanas and M. S. Kamel. Feature based decision fusion. In *Proceedings of Advances in Pattern Recognition - ICAPR 2001*, pages 176–185, Rio de Janeiro, Brazil, March 2001.

[36] I. Yaniv. Weighting and trimming: Heuristics for aggregating judgments under uncertainty. *Organizational Behavior and Human Decision Processes*, 69:237–249, 1997.

[37] L. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems 1(1)*, pages 3–28, 1978.