Journal of Global Information Management
Volume 28 • Issue 3 • July-September 2020

A Comparison of Retrieval Result Relevance Judgments Between American and Chinese Users

Jin Zhang, University of Wisconsin, Milwaukee, USA
https://orcid.org/0000-0002-6665-6606

Yuehua Zhao, Nanjing University, Nanjing, China
https://orcid.org/0000-0002-8412-2878

Xin Cai, University of Wisconsin, Milwaukee, USA

Taowen Le, Weber State University, Ogden, USA

Wei Fei, Suzhou Library, Suzhou, China

Feicheng Ma, Wuhan University, Wuhan, China
https://orcid.org/0000-0003-0187-0131

ABSTRACT

Relevance judgment plays an extremely significant role in information retrieval. This study investigates the differences between American users and Chinese users in relevance judgment during the information retrieval process. 384 sets of relevance scores with 50 scores in each set were collected from 16 American users and 16 Chinese users as they judged retrieval records from two major search engines based on 24 predefined search tasks from 4 domain categories. Statistical analyses reveal that there are significant differences between American assessors and Chinese assessors in relevance judgments. Significant gender differences also appear within both the American and the Chinese assessor groups. The study also revealed significant interactions among cultures, genders, and subject categories. These findings can enhance the understanding of cultural impact on information retrieval and can assist in the design of effective cross-language information retrieval systems.

KEYWORDS

Comparative Study, Factor Analysis, Information Management, Information Retrieval, Relevance Judgment, Search Engine, User Involvement

1. INTRODUCTION

As people throughout the world continue their reliance on the Internet to fulfill their information needs (Khatwani & Srivastava, 2017) and as the Internet continues its profound impact on people and societies (Teo, 2007; Lane, et al., 2017), researchers have explored ways to maximize successes of Internet-based technology implementations or global information management (Roztocki &

DOI: 10.4018/JGIM.2020070108

Weistroffer, 2011; Caprio, et al., 2015; Hung, et al., 2016; Silic & Back, 2016; Soja, 2016; Chatterjee, et al., 2017). Effective information retrieval from the Internet has remained an important topic in the field of global information management.

The key objective of information retrieval is to find relevant information. Therefore, the concept of relevance has been recognized as a fundamental issue to the evaluation of information retrieval systems as well as search engines (Borlund, 2003; Saracevic, 2007a; Zhang & Fei, 2010; Zhang et al., 2013). Relevance measures the effectiveness of a connection between a source and a target in a communication process (Saracevic, 1975).

Classic research on information retrieval concentrates on the techniques for the comparison of the representation of documents and search queries (Vakkari, 1999). Regarding the effectiveness of retrieval systems, traditional measures include precision and recall (Davidson, 1977). Typically, information retrieval systems treat all users the same way. However, more recent studies have regarded the information retrieval process as an interaction between users and the information retrieval system. From the user's perspective, information retrieval systems should be designed to satisfy user information needs.

Some believe that information and culture cannot be partitioned since culture is composed of different and transmitted social information (Kim, 2013). Culture has been demonstrated as a key factor for the understanding of human behavior because culture could influence perceptions, opinions, and ultimately reactions to behavior. Cultural differences may shape information-processing procedures (Gutchess & Indeck, 2009), and cross-cultural differences may affect individuals' information behaviors (Schwartz, et al, 2014). Some identified that information searching behavior differed between Chinese subjects and German subjects (Honold, 1999). Some explored differences in information seeking behaviors among software engineers from various countries and found that US and European subjects preferred non-social sources, while Indian and Pakistan participants favored social sources (Milewski, 2007). Therefore, understanding how cultural differences influence users' relevance judgment is crucial to information-retrieval research.

The primary research objective of this study is to explore possible differences between American users and Chinese users in relevance judgment of retrieval results from search engines. To achieve this purpose, relevance judgments by two assessor groups of retrieval result in different domain categories were compared, interactions between the assessor groups and the domain categories were analyzed, and gender differences in relevance judgments were explored. The findings of this study can enhance understanding of cultural differences in information retrieval and can assist in the design of effective cross-language search engines.

2. RELATED RESEARCH

2.1. Search Engine and Cross-Language Information Retrieval

According to the *Oxford English Dictionary* (2002), the search engine is "a piece of hardware or software designed for searching, *esp*. a program that searches for and identifies items in a database that correspond to one or more keywords specified by the user; *spec*. such a program used to search for information available over the Internet, using its own previously compiled database of Internet files and documents." This definition has emphasized users as an important component of search engines. However, Amichai-Hamburger (2002) asserts that there is a lack of awareness of the role of users in the design of Web systems and site content. Jansen and Spink (2006) have also demonstrated that there is a critical need to understand how people utilize Web search engines.

As the World Wide Web has been established as a global communication platform, information on the Internet is being produced in a variety of languages (Grefenstette, 2012). Although English has been adopted as the universal language, language barriers still hinder the broad information retrieval to some extent. In such situations, the fulfillment of the information needs of non-native

speakers becomes more common and significant. Cross-Language Information Retrieval (CLIR), as a branch of information retrieval, addresses the task of filtering, selecting, and ranking documents based on a query expressed in a different language (Grefenstette, 2012). In addition to the problems of monolingual information retrieval, different classes of approaches have been developed to deal with the translation problem: approaches using machine translation (MT) system, dictionary-based-translation approaches, and approaches based on parallel and comparable corpora (Nie, 2010).

2.2. Relevance Judgments

Discussions regarding relevance have primarily focused on two aspects: differentiation of various types of relevance, and definition of relevance in information retrieval (Maglaughlin & Sonnenwald, 2002). As viewed and applied in the context of information-retrieval research, relevance has been defined as a multidimensional and dynamic concept. Relevance is also divided into two general categories: system-oriented and user-oriented. The objective or system-based perspective treats relevance as a static and objective concept, whereas the subjective or human (user)-based approach considers relevance to be a subjective and personalized mental process that involves cognitive restructuring (Borlund, 2003).

This study adopted Wang and Soergel's definition of relevance: "relevance is a relationship between a need and a document judged by a person" (Wang & Soergel, 1998, p. 116). The degree of relevance refers to the rating of relevance between retrieval results and queries. In the evaluation of information retrieval systems, the assignments of relevance assessments could be binary relevance (relevant or non-relevant), partial relevance, scaled relevance, rated relevance, or three-valued or tri-partite relevance (Borlund, 2003).

The multidimensional nature of relevance can be illustrated by the different relevance criteria users employ to judge the relevance of search results (Borlund, 2003). A variety of studies obtained user-defined criteria directly from participants through think-aloud protocols, interviews, and questionnaires (Barry, 1998; Hirsh, 1999; Fitzgerald & Galloway, 2001; Maglaughlin & Sonnenwald, 2002). For example, through interviews, Park (1993) observed three major categories of variables affecting relevance assessments: internal context, external context, and problem context. Subsequently, Harter (1996) extracted 24 factors from Park (1993) and grouped them into four categories.

Relevance judgment is one of the most important steps within the information retrieval process, where time, context, and situation can influence the judgment decisions (Taylor, 2012). Relevance judgment is affected by a host of factors (Saracevic, 2007b). Previous research has shed light on possible variables concerning relevance judgments. In regard to cognitive processes, people differ in relevance-inference process, and thus affect their relevance decisions. Previous researchers have investigated individual differences in relevance judgment behaviors. Davidson (1977) examined that individual differences in openness to information may affect the relevance decisions. Domain knowledge also plays a role in searchers' relevance assessment. Dong, Loh, and Mondry (2005) divided 12 volunteers into two groups according to their domain knowledge. Then, all subjects were requested to search ten keywords in a search engine and assessed the relevance of retrieved objects. As a result, a higher degree of agreements existed among evaluators with more subject knowledge. In addition, Heinström (2003) explored the influence of five personality dimensions (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness) with users' information behavior. Heinström (2003) concluded that users' relevance judgment was related to a careless, competitive, sensitive, and conservative personality. Park (1993) concluded that people's experiences, perceptions, and private knowledge impacted their relevance assessments. Hansen and Karlgren (2005) recruited 28 participants whose first language was Swedish and who were fluent in English to search a newspaper database. They used several queries with results presented in Swedish and English to examine how assessors would evaluate the relevance of retrieved documents in a foreign language. Although extra efforts were invested into reading the English texts, Hansen and Karlgren (2005) discovered that the relevance assessment results were significantly less reliable for the English content than for the Swedish content.

2.3. Culture Difference

Previous research indicated that collectivism-individualism was the most significant dimension of national culture differences (Hofstede, 2001; Jackson & Wang, 2013). The US and China had been utilized as a pair of comparative examples. China was a prototypic example of eastern cultures, which were collectivistic, whereas the US was a prototypic example of western cultures, which were individualistic (Jackson & Wang, 2013).

Unique cultural characteristics may influence users' information behavior. Lee and Kwak (2016) conducted two independent surveys of adults in the US and South Korea to assess the extent to which cultures would shape the implications of mobile communication for deliberative democracy. Jackson and Wang (2013) surveyed 401 college students in China and 491 college students in the US and found that there were cultural differences between college students in China and US in the use of social networking sites, such as time spent, importance, and motives for use. In addition, by examining the motives for and patterns of using social network sites, Kim and associates asserted that cultural contexts shaped the use of communication technology among college students in the US and Korea (Kim, et. al., 2011).

Despite a rich literature on relevance judgments, few studies addressed the roles of culture and gender in user's relevance judgements or decisions. This study takes particular interest in the examination of the impact of culture in users' relevance judgments.

3. RESEARCH METHODOLOGY

In this study, the following three factors were considered: assessor groups, subject domain categories, and genders. The assessor group factor refers to the two groups of users or relevance assessors. One group consisted of 16 American participants while the other 16 Chinese participants. In total, there were 32 assessors in the study. The subject domain category factor refers to four subject categories of search tasks: Health, News and Media, Science and Technology, and Economy and Business. The gender factor refers to the gender of the relevance assessors: male and female.

There are many search engines on the Internet. Google and Bing were selected because they are widely-used top search engines. Statistical data showed that leading Internet search engines Google and Bing accounted for 89.95% and 3.99% of the global market shares in 2019, respectively, while the Chinese search engine Baidu accounted for only 0.56% of the market share (Statista, 2019).

Since it was reported that a sample size of about 30 is acceptable in an experimental study (Hogg and Tanis, 2005), 32 subjects were selected in this study. The subjects of the experiment were recruited through campus advertisement. The researchers advertised the research study and called for participation on the two campuses they attended in the United States and China respectively. Qualified American subjects and Chinese subjects were then selected for the study. To minimize interferences from user experience and expertise with the technology itself (Goeke, et al, 2016), all participants were college students as modern college students were typically familiar with Internet search systems. The 16 student evaluators in the American assessor group included 10 males and 6 females; they were students of Weber State University (WSU) in the U.S. The 16 student evaluators in the Chinese assessor group included 6 males and 10 females; they were students of Suzhou University (SZU) in China. All of them were proficient in English.

Since subject domain categories, genders, and assessor groups are the primary factors this study considered, to achieve the proposed research objective, we developed four hypothesis groups, targeting at these factors and their combinations. Group I explores relationships among the four domains, two assessor country groups, as well as their interactions in terms of relevance judgment; Group II discovers judgment differences pertaining to the male assessors; Group III examines judgment differences pertaining to the female assessors; Group IV reveals judgment differences between the American assessors and the Chinese assessors.

Group I: The first group of the hypotheses addressed two major variables: the subject domain and assessor group. In this hypothesis group, the performance of the relevance judgment pertaining to each of the two variables was examined. In addition, the interaction between the two variables was also tested:

- **($H1_0$):** There are no significant differences between the American assessors and the Chinese assessors in terms of relevance judgments;
- **($H2_0$):** There are no significant differences among the 4 subject domain categories in terms of relevance judgments;
- **($H3_0$):** There are no significant interactions between the assessor groups and the domain categories in terms of relevance judgments.

Group II: The second group of hypotheses focused on male performances. They ascertained the relevance-judgement differences in the male group between the two assessor groups:

- **($H4_0$):** There are no significant differences between the American male assessors and the Chinese male assessors in terms of relevance judgments;
- **($H5_0$):** There are no significant interactions between the male assessor groups and the domain categories in terms of relevance judgments.

Group III: The third group of hypotheses focused on female performances. They ascertained the relevance-judgement differences in the female group between the two assessor groups:

- **($H6_0$):** There are no significant differences between the American female assessors and the Chinese female assessors in terms of relevance judgments;
- **($H7_0$):** There are no significant interactions between the female assessor groups and the domain categories in terms of relevance judgments.

Group IV: The last group of hypotheses concentrated on relevance-judgement differences between males and females in three scenarios: (1) the American assessors and the Chinese assessors combined, (2) the Chinese assessors only; and (3) the American assessors only:

- **($H8_0$):** There are no significant differences between the male assessors and the female assessors in terms of relevance judgments;
- **($H9_0$):** There are no significant differences between the Chinese male assessors and the Chinese female assessors in terms of relevance judgments;
- **($H10_0$):** There are no significant differences between the American male assessors and the American female assessors in terms of relevance judgments.

The two factors for Group I, Group II, and Group III were assessor groups and subject domain categories. The dependent variable for the hypotheses was relevance judgments measured by the relevance scores assigned by relevance assessors. The relevance score was designed on an 11-points scale: 0 for totally irrelevant and 10 for most relevant.

3.1. Data Collection

Two popular search engines, Google and Bing, were used to produce retrieval result lists since they were widely regarded as the most popular search engines for Internet search (comScore, 2015). The topics of search tasks covered four general domain categories. The four categories used in this study were selected from the Yahoo! Answers subject directory with minor revisions (Yahoo! Answers, 2019). The Yahoo! Answers Website is a public question-and-answer forum. Its subject directory system is generated based on users' information needs. In other words, the Yahoo categories are user-oriented. The principle of the category selection is to include categories general enough to fit the participants of both countries and to exclude categories too specific such as Yahoo Products, Local Business, Dining Out, and Pregnancy and Parenting. Within each category, 6 search tasks were carefully generated to represent the category and used for data collection. For instance, topics such as *Lady Gaga*, *Obamacare, Bin Laden death, The Korean crisis*, *The Syria crisis*, and *H7N9 bird*

flu were included in the category of News and Media. Table 1 summarizes the 4 categories and the search tasks in each category.

In Table 1, the string in parentheses behind a category or task represents a specific search task for a specific category. For instance, *C1_T1* stands for the search task of Autism in the category of

Table 1. Summary of subject domain categories and search tasks

Health (C1)	News and Media (C2)	Science and Technology (C3)	Economy and Business (C4)
Autism (C1_T1)	Lady Gaga (C2_T1)	Google glasses (C3_T1)	BRICS (C4_T1)
Weight control (C1_T2)	Obamacare (C2_T2)	Global warming and climate change (C3_T2)	World Trade Organization (C4_T2)
Smoking and health (C1_T3)	Bin Laden death (C2_T3)	Web 2.0 (C3_T3)	US dollar and Chinese Yuan exchange rate (C4_T3)
AIDS prevention (C1_T4)	The Korean crisis (C2_T4)	Wind energy (C3_T4)	Hedge Fund (C4_T4)
Asthma (C1_T5)	The Syria crisis (C2_T5)	Electric car (C3_T5)	The Big Mac Index (C4_T5)
Birth control (C1_T6)	H7N9 bird flu (C2_T6)	Stem cell research (C3_T6)	Micro-economy (C4_T6)

Health. Each search task corresponded to a search query. Each query was submitted to both Google and Bing.

A total of 48 (4 domain categories, 6 search tasks from each category, and each search task submitted to 2 search engines, hence 4 x 6 x 2 = 48) retrieval result lists were produced with each result list consisting of 50 records.

3.2. Relevance Judgment

Relevance is an important dependent variable in this study. Its measurement would affect the validity of the study findings. While the famous TREC datasets were used in many traditional information-retrieval-evaluation studies (TREC, 2019), item relevance in the databases was only on a 2-point scale. Relevance measurement of finer granularity would lead to more reliable results because subtle differences of relevance-judgment results can be affectively detected and considered in data analysis. It is crucial for the retrieval-results ranking analysis in a search engine. On the other hand, Nunnally and Bernstein (1994) suggest in their psychometric research study that although a Likert scale design with more scale points is better than a design with less scale points, there is a diminishing return after 11 points. As a result, an 11-point scale was selected for the relevance measurement in this study.

After all retrieval result lists were produced, the next step was relevance judgment for the records in each of the result lists. 32 assessors (16 American and 16 Chinese) participated in the relevance-judgment process.

The 32 assessors were divided into two groups: the group of 16 American accessors and the group of 16 Chinese accessors. 8 American accessors and 8 Chinese accessors were randomly assigned to judge the relevance of the Google retrieval results; Out of them, 4 American accessors and 4 Chinese accessors were randomly assigned to evaluate retrieval results of the first 12 search tasks, while the other 4 American accessors and the other 4 Chinese accessors were randomly assigned to evaluate retrieval results of the remaining 12 search tasks. Likewise, 8 American accessors and 8 Chinese

accessors were randomly assigned to evaluate the relevance of the Bing retrieval results; Out of them, 4 American accessors and 4 Chinese accessors were randomly assigned to evaluate retrieval results of the first 12 search tasks, and the other 4 American accessors and the other 4 Chinese accessors were randomly assigned to evaluate retrieval results of the remaining 12 search tasks.

Each assessor was required to read a search task and understand its meaning thoroughly before making decision regarding relevance of the retrieved records. If they had questions on any search task, they were to ask the researchers for clarification. They would read the title and the content of a retrieved record and then assign a relevance score for the record.

In summary, each of the 48 retrieval result lists (with 50 records in each list) were evaluated by 8 accessors (4 from the U.S. and 4 from China), yielding a total of 384 (48 x 8 = 384) sets of relevance scores, with each set containing 50 records and therefore 50 individual scores. The same could also be stated as each of the 32 accessors evaluating 12 retrieved result lists and producing 384 (32 x 12 = 384) sets of relevance scores.

3.3. Hypothesis Testing

The two-factor *ANOVA* tests were conducted to test the proposed null hypotheses in Groups I, II, and III. The t-test approaches were used to test the null hypotheses in Group IV. The significance level of 0.05 was used for hypothesis rejection or not. In other words, if the *p*-value produced from an inferential test were larger than 0.05, the corresponding null hypothesis would not be rejected; otherwise, the hypothesis would be rejected. If the hypothesis is rejected, a follow-up *Tukey* test will be conducted to detect which assessor groups and/or domain categories resulted in the rejection. The statistics software of *SPSS* (Version 20; IBM Corp, 2011) was used for the hypothesis testing.

4. RESULTS AND DISCUSSIONS

4.1. The Descriptive Summary

In this study, two search engines (Google and Bing) were used to retrieve Webpages for relevance judgment. 6 queries (or search tasks) from each of 4 domain categories were submitted to each of the two search engines, and the first 50 retrieval records from each query were captured and saved for relevance judgment. Consequently, $6 \times 4 \times 2 \times 50 = 2400$ Webpages were captured and collected for relevance judgment. Each Webpage was examined by 8 evaluators (4 Americans and 4 Chinese). Therefore, a total of 19200 $(2400 \times 8 = 19200)$ individual relevance scores were collected in this study.

Figure 1 illustrates the distribution of the relevance score frequencies. In Figure 1, the *X*-axis represents relevance scores while the *Y*-axis represents the frequency of the relevance scores. The least occurring relevance score was 0, which only appeared 7 times. The most-frequently occurring score was 8 (3082 times), which accounted for 19.80% of all data.

Figure 2 shows the descriptive summary of the relevance scores. In this figure, the American assessor group refers to the WSU students while the Chinese assessor group refers to the SZU students. The mean of the relevance scores was 6.9373, and the standard deviation was 2.1968. The mean of the Chinese assessor group was 6.6547, and its standard deviation was 1.73887. For the Chinese assessor group, the respective means of the four categories were in descending order News and Media (C2) (6.7562), Science and Technology (C3) (6.7325), Economy and Business (C4) (6.6488), and Health (C1) (6.6417). The mean of the American assessor group (7.2199) was higher than that of the Chinese assessors, and the standard deviation (2.2491) was also higher. For the American assessor group, the category with the lowest mean was News and Media (C2) (7.0654), while the same category had the highest category mean for the Chinese assessor group (6.7562). The category means of the other three categories for the American accessor group were in ascending order Economy and Business (C4) (7.2142), Health (C1) (7.2946), and Science and Technology (C3) (7.3054).

Figure 1. Relevance score frequencies

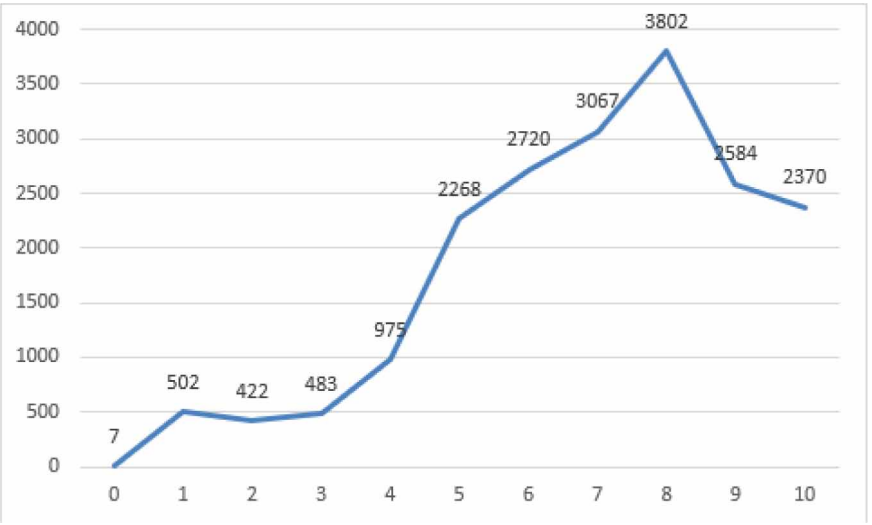


Figure 2. Relevance scores of American assessor group and Chinese assessor group in the 4 domain categories

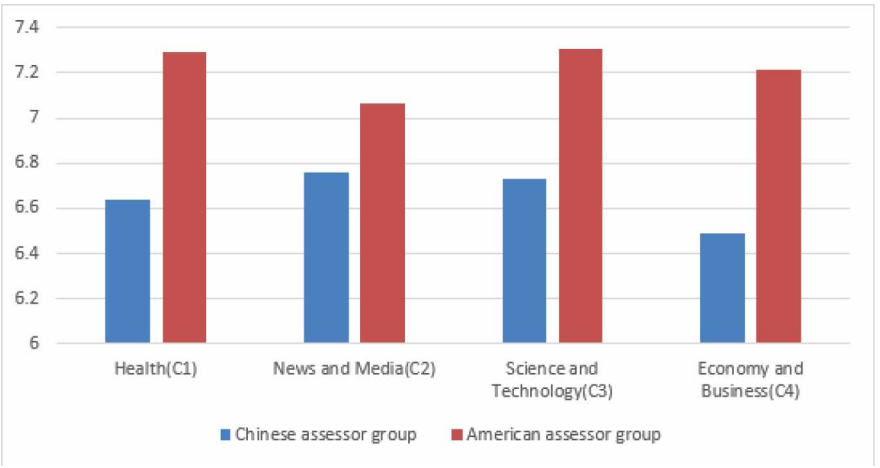
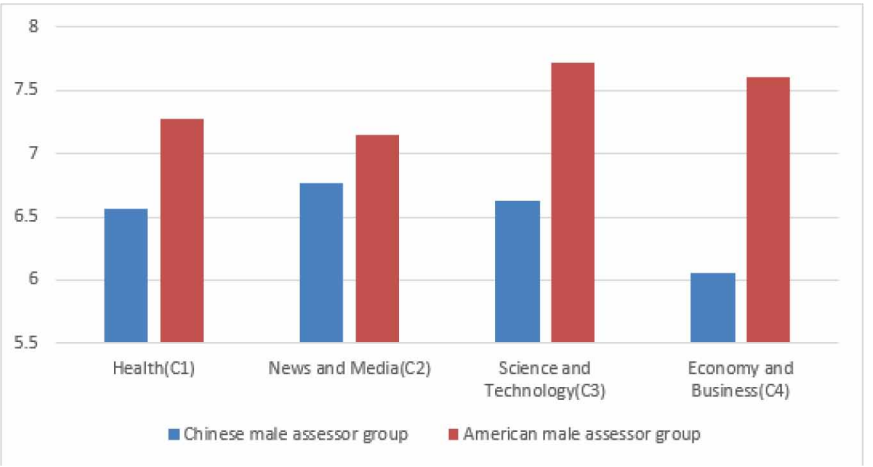


Figure 3 compares the mean relevance scores assigned by the American male assessors and those assigned by the Chinese male assessors. The mean relevance score for the combined male assessors was 7.1534, and the standard deviation was 2.1300. The Chinese male assessors assigned an average relevance score of 6.6083, and the standard deviation was 1.6486. In respect to domain categories, the Chinese male assessors considered the retrieved results in News and Media (C2) (6.7627) more relevant than in other categories, whereas the standard deviation was the lowest (1.5375). The means of the other categories were in descending order Science and Technology (C3) (6.6333), Health (C1) (6.56), and Economy and Business (C4) (6.0533). The mean of the American male assessor group was 7.4805, while the standard deviation was 2.31168. Unlike their Chinese peers, the American male assessor considered the retrieved results in Science and Technology (C3) (7.7211) to be more relevant than in other categories, and the standard deviation was 2.1277. The means of the other

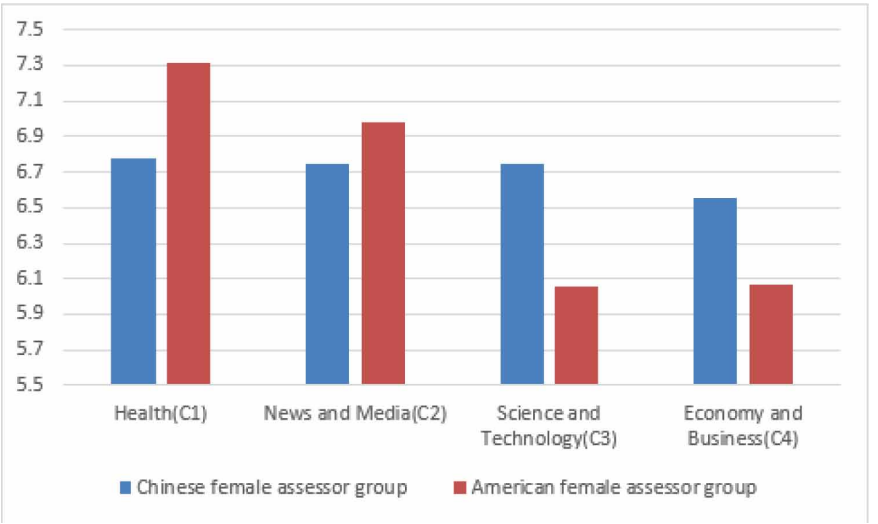
Figure 3. Relevance scores of American male assessors and Chinese male assessors in the 4 domain categories



categories were in an ascending order News and Media (C2) (7.1475), Health (C1) (7.2783), and Economy and Business (C4) (7.5967).

Figure 4 shows the mean relevance scores of the American female assessor group and the Chinese female assessor group. The combined female assessors group assigned an average relevance score of

Figure 4. Relevance scores of the American female assessors and the Chinese female assessors in the 4 domain categories



6.7211 with a standard deviation of 2.2409. The mean for the Chinese female assessors was 6.6825 with a standard deviation of 1.7904. The Chinese female assessors considered the retrieved results from *Health* (C1) (6.7778) more relevant than in other categories. The means of the other categories were in ascending order *Economy and Business* (C4) (6.5505), *News and Media* (C2) (6.7456), and *Science and Technology* (C3) (6.7467). Interestingly, the American female assessors also regarded the retrieval results from *Health* (C1) (7.3108) as most relevant. The means of the other categories

were in descending order *News and Media* (C2) (6.9833), *Economy and Business* (C4) (6.0667), and *Science and Technology* (C3) (6.0583). The mean relevance score assigned by American female assessors (6.7856) was higher than that assigned by their counterparts, and the standard deviation (2.8362) was also higher.

4.2. Inferential Statistics Analysis

4.2.1. Results for Hypotheses $H1_0$, $H2_0$, and $H3_0$

A two-factor *ANOVA* test was conducted to examine Hypotheses $H1_0$, $H2_0$, and $H3_0$. Hypothesis $H1_0$ *was* proposed to examine the difference between the American assessor group and the Chinese assessor group in terms of relevance judgments. Hypothesis $H2_0$ was to investigate the differences among the 4 defined domain categories. Hypothesis $H3_0$ was to detect potential interactions between the assessor group factor and the domain category factor.

Table 2 summarizes the test results for $H1_0$, $H2_0$, and $H3_0$. For Hypothesis $H1_0$, with *df* (1, 19192), the *F* value (323.666) is larger than the critical value (3.84) at the significant level (0.05).

Table 2. Test results for $H1_0$, $H2_0$, and $H3_0$

Factor	Type III Sum of Squares	df	Mean Square	F	p Value
Assessor Group	1533.41	1	1533.41	323.666	0.000
Category	75.472	3	25.157	5.310	0.001
Interactions	118.932	3	39.644	8.368	0.000

The *p* value (0.000) is smaller than the significant level (0.05). As a result, $H1_0$ is rejected, and there are significant differences between the American assessor group and the Chinese assessor group in terms of their relevance judgments of retrieved results from search engines.

For Hypothesis $H2_0$, with *df* (3, 19192), the *F* value (5.310) is larger than the critical value (2.61) at significant level (0.05). The *p* value (0.001) is smaller than the significant level (0.05). Therefore, $H2_0$ is rejected, and there are significant differences among the 4 defined domain categories in terms of relevance judgments of retrieved results from search engines.

For Hypothesis $H3_0$, with *df* (3, 19192), the critical value at significant level (0.05) is 2.61, which is smaller than the *F* value (8.368), and the *p* value (0.000) is smaller than the significant level (0.05). Therefore, $H3_0$ is rejected, and there are significant interactions between the assessor group factor and the domain category factor in terms of relevance judgments of retrieved results from search engines.

A follow-up Tukey test was conducted to detect reasons of the rejection of $H2_0$. Table 3 shows the results. In this Table, I and J stand for categories, and I-J is the mean difference of the relevance scores between Category I and Category J. For instance, mean difference (1-2) is the mean difference between categories 1 and 2 in terms of relevance scores. A mean difference with an asterisk indicates that the difference is significant between the two categories. Categories 1 to 4 represent Health, News and Media, Science and Technology, and Economy and Business respectively. Table 3 shows that the hypothesis rejection was caused by the significant differences between Health and Economy and Business (0.1169) and between Science and Technology and Economy and Business (-0.1677).

Table 4 shows the results of grouping domain categories in terms of relevance scores. It suggests that there are 2 homogeneous groups. Group 1 contains domain categories News and Media (C2) and Economy and Business (C4), while group 2 includes Health (C1), News and Media (C2), and Science and Technology (C3).

Table 3. Follow-up Tukey test results for $H2_0$

(I) Category	(J) Category	Mean Difference (I-J)	Std. Error	p Value
1	2	0.0573	0.04443	0.57
	3	-0.0508	0.04443	0.662
	4	.1169*	0.04443	0.042
2	1	-0.0573	0.04443	0.57
	3	-0.1081	0.04443	0.071
	4	0.0596	0.04443	0.537
3	1	0.0508	0.04443	0.662
	2	0.1081	0.04443	0.071
	4	.1677*	0.04443	0.001
4	1	-.1169*	0.04443	0.042
	2	-0.0596	0.04443	0.537
	3	-.1677*	0.04443	0.001

Table 4. Results of grouping domain categories for $H2_0$

Category	N	Subset	
		1	2
4	4800	6.8512	
2	4800	6.9108	6.9108
1	4800		6.9681
3	4800		7.019
p value		0.537	0.071

The rejection of $H3_0$ indicates that there are significant interactions between the assessor group factor and the domain category factor. Figure 5 shows the interactions of the two factors. In Figure 5, the *X*-axis represents the assessor groups and the *Y*-axis represents the mean relevance scores. Each line represents the means of the two assessor groups for a specific domain category. It is obvious that News and Media (C2) has intersections with all other three domain categories. It means that the rejection of $H3_0$ is mainly caused by News and Media (C2).

Another two-factor *ANOVA* was conducted to test Hypotheses $H4_0$ and $H5_0$. Hypothesis $H4_0$ was proposed to investigate the difference between the American male assessors and the Chinese male assessors in terms of relevance judgments. Hypothesis $H5_0$ was proposed to examine interactions between the male assessor groups and the domain categories.

Table 5 summarizes the test results for $H4_0$ and $H5_0$. For $H4_0$, with *df* (1, 9592), the *F* value (299.795) is larger than the critical value (3.84) at signification level (0.05), and the *p* value (0.000) is smaller than the significant level (0.05). It suggests that Hypothesis $H4_0$ is rejected, and there are significant differences between American male assessors and Chinese male assessors in terms of relevance judgments of retrieved results from search engines.

For Hypothesis $H5_0$, with *df* (3, 9592), the critical value at significant level (0.05) is 2.61, which is smaller than the *F* value 21.655, and the *p* value (0.000) is also smaller than the significant level

Figure 5. Interactions between the assessor group factor and the domain category factor results for hypotheses $H4_0$ and $H5_0$

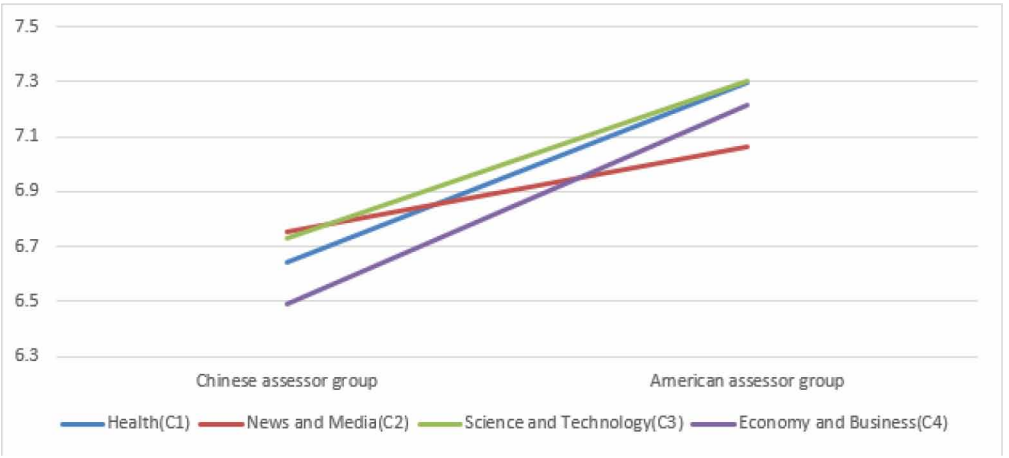


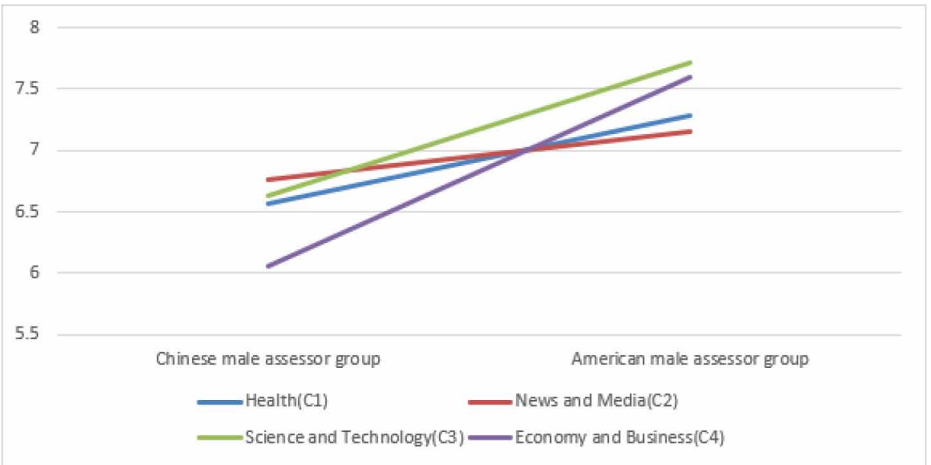
Table 5. Test results for $H4_0$ and $H5_0$

Factor	Type III Sum of Squares	df	Mean Square	F	p Value
Assessor Group	1293.850	1	1293.850	299.795	0.000
Category	71.601	3	23.867	5.530	0.001
Interactions	280.370	3	93.457	21.655	0.000

(0.05). Therefore, $H5_0$ is rejected, and there are significant interactions between the male assessor groups and the domain categories.

Since $H5_0$ is rejected, Figure 6 is drawn to detect the reasons of the rejection. In Figure 6, the *X*-axis represents the male assessor groups, and the *Y*-axis represents the mean relevance scores. It is apparent that relevance judgments differ the most between the American male assessor group and the

Figure 6. Interactions between the male assessor groups and the domain categories results for hypotheses $H6_0$ and $H7_0$



Chinese male assessor group in Economy and Business (C4) and the least in News and Media (C2). It indicates that the rejection of Hypothesis $H5_0$ is mainly caused by these two domain categories.

The last two-factor *ANOVA* was performed to test Hypotheses $H6_0$ *and* $H7_0$. Hypothesis $H6_0$ was proposed to investigate the difference between the American female assessors and the Chinese female assessors in terms of their relevance judgments. Hypothesis $H7_0$ was to examine interactions between the female assessor groups and the domain categories.

Table 6 summarizes the test results. For Hypothesis $H6_0$, with *df* (1, 9592), the *F* value of $H6_0$ (4.011) is larger than the critical value (3.84) at significant value (0.05), and the *p* value (0.045) is

Table 6. Test results for $H6_0$ and $H7_0$

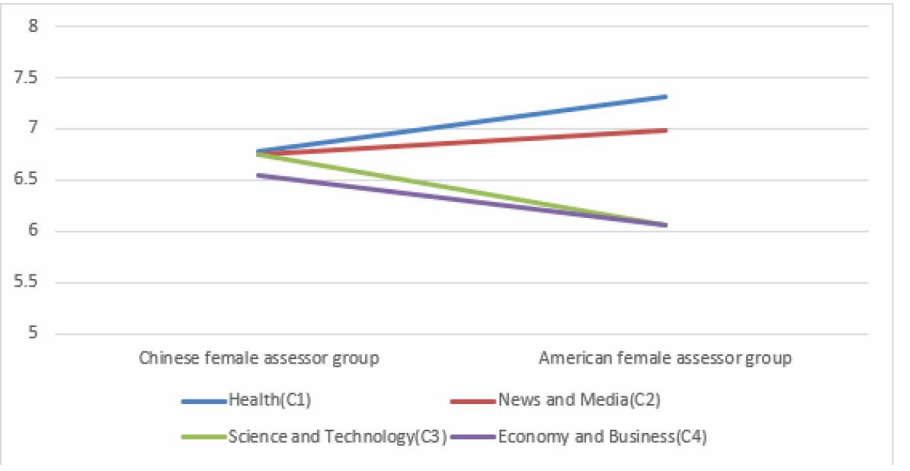
Factor	Type III Sum of Squares	df	Mean Square	F	p Value
Assessor Group	19.701	1	19.701	4.011	0.045
Category	743.406	3	247.802	50.446	0.000
Interactions	493.996	3	164.665	33.521	0.000

smaller than the significant value (0.05). It signifies that Hypothesis $H6_0$ is rejected, and there are significant differences between the American female accessors and the Chinese female accessors in terms of their relevance judgments.

For Hypothesis $H7_0$, with *df* (3, 9592), the *F* value (33.521) is larger than the critical value (2.61) at significant value (0.05), while the *p* value (0.000) is smaller than the significant level (0.05), which suggests that Hypothesis $H7_0$ is rejected, and there are significant interactions between the assessor groups and the domain categories in terms of relevance judgments of retrieved results from search engines.

In order to uncover the reasons of the rejection of Hypothesis $H7_0$, Figure 7 is drawn to show the interactions of the variables. In Figure 7, the *X*-axis represents the assessor groups, and the *Y*-axis represents the mean relevance scores. It is observed that while the Chinese female assessors have similar means of relevance judgments in the 4 domain categories, the American female assessors

Figure 7. Interactions between the female assessor groups and the domain categories results for hypotheses $H8_0$, $H9_0$ and $H10_0$



judged retrieval results in Science and Technology (C3) and Economy and Business (C4) to be less relevant than those in Health (C1) and News and Media (C2). These facts contributed to the rejection of Hypothesis $H7_0$.

Three *t*-test analyses were conducted to test Hypotheses $H8_0$, $H9_0$ and $H10_0$. Hypothesis $H8_0$ was proposed to examine the difference between the male assessor group and the female assessor group in terms of relevance judgment. Table 7 shows the test results for $H8_0$. With *df* (19198), the *t* value (13.7)

Table 7. Test results for $H8_0$

t Value	Df	Mean Difference	p Value (2-Tailed)
13.7	19198	0.432	0.000

is larger than the critical value (1.96) at significant level (0.05), and the *p* value (0.000) is smaller than the significant value (0.05). It suggests that $H8_0$ is rejected, and there are significant differences between the male assessor group and the female assessor group in terms of relevance judgments.

Hypothesis $H9_0$ was proposed to examine the difference between the Chinese male assessors and the Chinese female assessors. Table 8 shows the results. With df (9598), the *t* value (-2.023) is

Table 8. Test results for $H9_0$

t Value	df	Mean Difference	p Value
-2.023	9598	-0.074	0.043

smaller than the critical value (-1.96) at significant level (0.05), and the p value (0.043) is smaller than the significant value (0.05). It suggests that the $H9_0$ is rejected, and there are significant differences between the Chinese male assessors and the Chinese female assessors in their relevance judgments.

The last hypothesis $H10_0$ was proposed to examine the difference between the American male assessors and the American female assessors in regard to their relevance judgments. Table 9

Table 9. Test results for $H10_0$

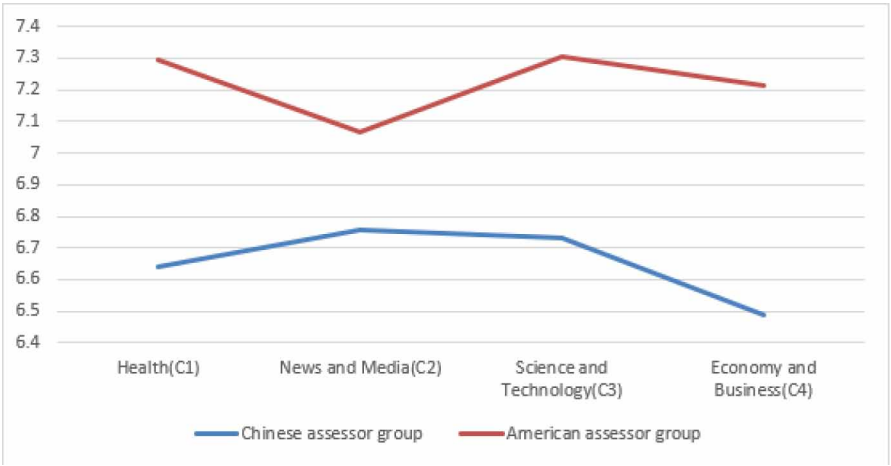
t Value	df	Mean Difference	p Value
13.075	9598	0.695	0.000

summarizes the results. With *df* (9598), the *t* value (13.075) is smaller than the critical value (1.96) at significant level (0.05), and the p value (0.000) is smaller than the significant value (0.05). It indicates that $H10_0$ is rejected, and there are significant differences between the American male assessors and the American female assessors in their relevance judgements.

5. DISCUSSION

In general, the American assessors considered the retrieved results more relevant than the Chinese assessors in all 4 defined domain categories. Figure 8 depicts the mean relevance scores of the

Figure 8. Mean relevance scores of American assessors and Chinese assessors in the 4 domain categories



American assessor group and the Chinese assessor group in the 4 domain categories. In Figure 8, the *X*-axis represents the domain categories, and the *Y*-axis represents the mean relevance scores. It shows that the mean relevance score of the American assessor group (7.2199) is significantly larger than that of the Chinese assessor group (6.6547) regardless of the domain categories. Two factors may have resulted in this difference: language and culture.

Although all Chinese accessors were proficient in English, English was not their native language. Since all search tasks, search engines, and retrieved Webpages were in English, a possible lack of thorough understanding of webpage content could have led to lower relevance score assignments. However, the same could have also led to higher relevance score assignments.

A more likely cause would be cultural differences. The Chinese culture has been known for its conservatism, while the American culture which finds its roots in Europe has been known for its liberalism. Such cultural difference could easily play a role in score assignments. An additional culture-related consideration would be the cultural nature of certain topics. For instance, the subject of "Obamacare" from News and Media (C2) was familiar to Americans but hardly known to the Chinese. A lack of cultural understanding could have contributed to the difference in relevance score assignments also.

Although the mean relevance scores of American assessors were higher than those of the Chinese assessors, the differences did not evenly exist across the 4 subject domain categories. The category of News and Media received the highest mean relevance score from the Chinese assessors (6.7562) but the lowest mean relevance score from the American assessors (7.0654). This might be explainable by examining the query terms (or search tasks) in Table 1. For example, the query "Lady Gaga" in News and Media produced vastly retrieval records containing "Lady Gaga" in their titles. Judging by the titles, the Chinese assessors who were not thoroughly familiar with the cultural content of the subject might think they were highly relevant; however, the American assessors who were more familiar with the related culture might find some not so relevant.

The Chinese assessor group also had a relative higher mean relevance score in Science and Technology (C3) (6.7325).

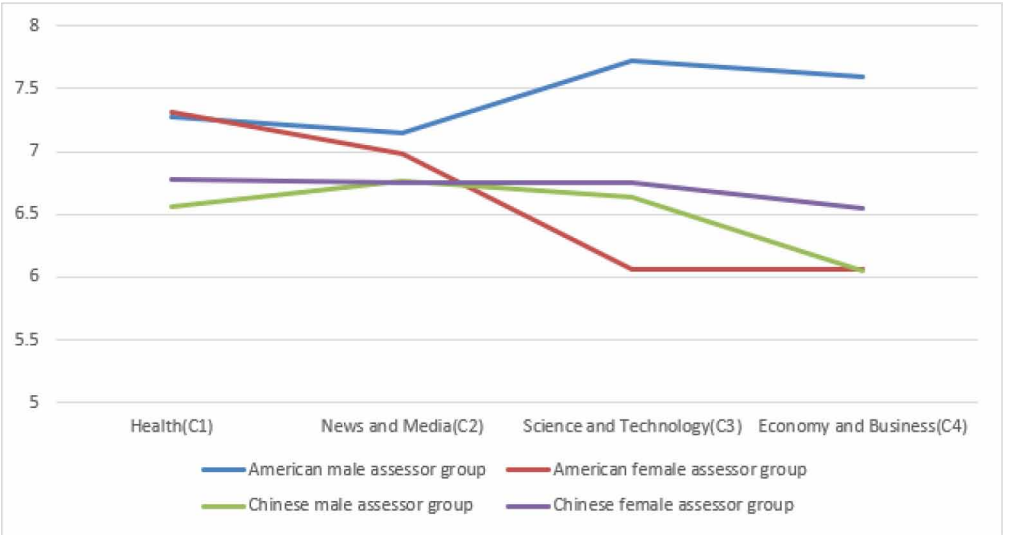
The largest difference (0.7259) between the Chinese assessors and the American assessors occurred in Economy and Business (C4). This could have been caused by insufficient familiarity with certain query terms. For example, the Chinese assessors might know the association of five major emerging national economies (Brazil, Russia, India, China and South Africa), but they might

not realize that "BRICS" is the acronym coined for that association. On the other hand, the American accessors might have no difficulties with either.

The Chinese assessors also rated the Health (C1) domain relatively lower (6.6417). This could have been caused by cultural factors. For instance, an average Chinese person was likely unaware of "autism", let alone considering autism as a severe mental disease that would need to be treated.

Significant gender differences in relevance judgments were observed within both the American assessor group and the Chinese assessor group. However, the differences between American male assessors and American female assessors were larger than the differences between Chinese male assessors and Chinese female assessors. Figure 9 depicts the mean relevance scores of the American male assessors, the Chinese male assessors, the American female assessors, and the Chinese female assessors in the 4 domain categories. In Figure 9, the X-axis represents the domain categories, and

Figure 9. Mean relevance scores of American male assessors, Chinese male assessors, American female assessors, and Chinese female assessors in 4 domain categories



the Y-axis represents the mean relevance scores. It is clear that while the curve of the Chinese male assessors and the curve of the Chinese female assessors were similar to each other, the curves of the American male assessors and the American female assessors did not follow the same patterns. The significant differences between the American male assessors and the American female assessors were caused mainly by the categories of Science and Technology (C3) and Economy and Business (C4).

Both the American participants and the Chinese participants in this study were undergraduate students. The majors of the American students were business, engineering, computer science, communications, psychology, and graphic design while the Chinese counterparts all majored in library and information science. Therefore, the American participants were more diverse than the Chinese participants. Such a difference in background could have affected the participants' relevance judgments.

In reality, information relevance judgment depends on users' information needs while the information needs are dynamic, diverse, and complicated. Relevance judgment is a subjective, situational, and cognitive process. It is determined by timeliness, novelty, authority, completeness, usefulness, and other related factors of an assessed item. In this study, the researchers did not ask the participants the reason why an item was selected as relevant or irrelevant. As a result, it is impossible to decide whether the relevance judgment was made based on the about-ness or the of-ness of the item.

Since each search engine has its own databases, responses to the same search query vary from different search engines. If different search engines are used, they result in different retrieval items. Therefore, the relevance-judgment results of the retrieved items could be different. Even with the same search engine and the same search query, responses to the query may be different if the query is submitted at different times, because the databases of the search engine are constantly updated, with new information constantly added to the databases. However, as long as the experimental procedure remains the same, the relevance-judgment results from different search engines (or the same search engine with different responses) should not be significantly different. In other words, the findings of the study based on different search engines can be generalized.

5.1. Implications

The findings of this study have both practical and theoretical implications. The differences in relevance judgments between American accessors and Chinese accessors may have significant implications for the design of cross-language search engines and information retrieval systems. The differences in relevance judgments among various domain categories may validate the need for a background support mechanism built into these systems to provide end-users with more background information regarding certain topics, such as those in science and technology.

The study has confirmed that there are significant differences in relevance judgment between users of different cultures. The findings not only help researchers and practitioners in information science better understand relevance-judgment behavior of users of different cultures, but also assist designers of search engines and information retrieval systems in developing culture-friendly systems. For instance, when Google, Bing, and other similar search engines process queries from non-English countries, each item on a retrieval-results list could include some background information and a translation version of the item in addition to regular content. It would help searchers make proper decisions in relevance judgment of the item.

Previous studies have identified a number of factors that might affect users' relevance judgment, including openness to information (Davidson, 1977), domain knowledge (Dong, Loh, and Mondry, 2005), personality (Heinström, 2003), and verbal comprehension skill (Samimi, Ravana, & Koh, 2016). Theoretically, this study suggests that culture differences, subject categories, and gender differences all influence how users assess the relevance of retrieved results.

6. CONCLUSION

Information retrieval is one of the most essential topics in the field of information science or information technology. Discovering and studying factors and variables affecting assessors' relevance judgment provides a lens into understanding the information retrieval process and information seeking behavior. Prior research has identified several crucial variables that affect information searchers' assessments of relevance. The current study investigates the differences between Chinese assessors and American assessors in terms of the relevance judgments of retrieval results from search engines. This study adds to the body of knowledge regarding relevance judgments with findings that culture differences, subject categories, and gender differences influence assessors' relevance assessment in the information retrieval process.

The study employed 2 user groups, 16 American participants and 16 Chinese participants, to evaluate relevance of records retrieved from search engines. The study utilized 24 search tasks covering the following 4 domain categories: Health, News and Media, Science and Technology, and Economy and Business. Two primary search engines, Google and Bing, were used to produce retrieval records for evaluation. The participants were asked to assign a relevance score for each retrieval record associated with each search task. Inferential statistical analyses were conducted to examine differences in assessors' relevance judgments.

One of the major findings of this study is that there are significant differences between American assessors and Chinese assessors in their relevance judgments. The American assessors assigned significantly higher relevance scores than the Chinese assessors in every domain category, although the difference was smaller in the *News and Media* category. Significant gender differences were discovered within both the American assessor group and Chinese assessor group. While the Chinese female assessors assigned higher relevance scores than the Chinese male assessors in all defined domain categories except *News and Media*, the American female assessors assigned lower relevance scores than the American male assessors in all defined domain categories except *Health*. The disagreements of relevance assessments were more apparent between the American male assessors and the American female assessors than between the Chinese male assessors and the Chinese female assessors.

The limitations of this study are: (1) only the search results extracted from the search engines of Google and Bing and in four domain areas were used in the experiment. If more search engines and more search topics were investigated, the results obtained would be more reliable; (2) all the participants in this study were college students. If the experiment could include other user groups such as high school students and senior citizens, the findings attained would be more generalizable. One of the future research directions is to investigate through qualitative research methods (such as interviews or think-aloud protocols) the reasons why certain user groups assigned higher or lower relevance scores in certain domain areas. There are other factors that could have affected the relevance judgment of the participants such as the participants' majors. The impact of the participants' majors on relevance judgment can be another research direction in the future.

REFERENCES

Amichai-Hamburger, Y. (2002). Internet and personality. *Computers in Human Behavior*, *18*(1), 1–10. doi:10.1016/S0747-5632(01)00034-6

Barry, C. L. (1998). Document representations and clues to document relevance. *Journal of the American Society for Information Science*, *49*(14), 1293–1303. doi:10.1002/(SICI)1097-4571(1998)49:14<1293::AID-ASI7>3.0.CO;2-E

Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, *54*(10), 913–925. doi:10.1002/asi.10286

Caprio, D. D., Santos-Arteaga, F. J., & Tavana, M. (2015). Technology Development through Knowledge Assimilation and Innovation: A European Perspective. *Journal of Global Information Management*, *23*(2), 48–93. doi:10.4018/JGIM.2015040103

Chatterjee, S., Kar, A. K., & Gupta, M. P. (2017). Critical Success Factors to Establish 5G Network in Smart Cities: Inputs for Security and Privacy. *Journal of Global Information Management*, *25*(2), 15–37. doi:10.4018/JGIM.2017040102

comScore. (2015). comScore Releases September 2015 U.S. Desktop Search Engine Rankings. Retrieved from http://www.comscore.com/Insights/Rankings/comScore-Releases-September-2015-US-Desktop-Search-Engine-Rankings

Davidson, D. (1977). The effect of individual differences of cognitive style on judgments of document relevance. *Journal of the American Society for Information Science*, *28*(5), 273–284. doi:10.1002/asi.4630280507

Dong, P., Loh, M., & Mondry, A. (2005). Relevance similarity: An alternative means to monitor information retrieval systems. *Biomedical Digital Libraries*, *2*(1), 6. doi:10.1186/1742-5581-2-6 PMID:16029513

Fitzgerald, M. A., & Galloway, C. (2001). Relevance judging, evaluation, and decision making in virtual libraries: A descriptive study. *Journal of the American Society for Information Science and Technology*, *52*(12), 989–1010. doi:10.1002/asi.1152

Goeke, R. J., Faley, R. H., Brandyberry, A. A., & Dow, K. E. (2016). How Experience and Expertise Affect the Use of a Complex Technology. *Information Resources Management Journal*, *29*(2), 59–80. doi:10.4018/IRMJ.2016040104

Grefenstette, G. (2012). *Cross-Language Information Retrieval*. Springer Science & Business Media.

Gutchess, A. H., & Indeck, A. (2009). Cultural influences on memory. *Progress in Brain Research*, *178*, 137–150. doi:10.1016/S0079-6123(09)17809-3 PMID:19874966

Hansen, P., & Karlgren, J. (2005). Effects of foreign language and task scenario on relevance assessment. *The Journal of Documentation*, *61*(5), 623–639. doi:10.1108/00220410510625831

Harter, S. P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, *47*(1), 37–49. doi:10.1002/(SICI)1097-4571(199601)47:1<37::AID-ASI4>3.0.CO;2-3

Heinström, J. (2003). Five personality dimensions and their influence on information behaviour. *Information Research*, *9*(1), 9–1.

Hirsh, S. G. (1999). Children's relevance criteria and information seeking on electronic resources. *Journal of the American Society for Information Science*, *50*(14), 1265–1283. doi:10.1002/(SICI)1097-4571(1999)50:14<1265::AID-ASI2>3.0.CO;2-E

Hofstede, G. (2001). Culture's Recent Consequences: Using Dimension Scores in Theory and Research. *International Journal of Cross Cultural Management*, *1*(1), 11–17. doi:10.1177/147059580111002

Hogg, R. V., & Tanis, E. A. (2005). *Probability and Statistical Inference*. Upper Saddle River, NJ: Prentice Hall.

Honold, P. (1999). Learning How to Use a Cellular Phone: Comparison between German and Chinese Users. *Technical Communication: Journal of the Society for Technical Communication*, *46*(2), 196–205.

Hung, S. Y., Huang, W. M., Yen, D. C., Chang, S. I., & Lu, C. C. (2016). Effect of Information Service Competence and Contextual Factors on the Effectiveness of Strategic Information Systems Planning in Hospitals. *Journal of Global Information Management*, *24*(1), 14–36. doi:10.4018/JGIM.2016010102

Jackson, L. A., & Wang, J.-L. (2013). Cultural differences in social networking site use: A comparative study of China and the United States. *Computers in Human Behavior*, *29*(3), 910–921. doi:10.1016/j.chb.2012.11.024

Jansen, B. J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management: An International Journal*, *42*(1), 248–263. doi:10.1016/j.ipm.2004.10.007

Khatwani, G., & Srivastava, P. R. (2017). An Optimization Model for Mapping Organization and Consumer Preferences for Internet Information Channels. *Journal of Global Information Management*, *25*(2), 88–115. doi:10.4018/JGIM.2017040106

Kim, J.-H. (2013). Information and culture: Cultural differences in the perception and recall of information. *Library & Information Science Research*, *35*(3), 241–250. doi:10.1016/j.lisr.2013.04.001

Kim, Y., Sohn, D., & Choi, S. M. (2011). Cultural difference in motivations for using social network sites: A comparative study of American and Korean college students. *Computers in Human Behavior*, *27*(1), 365–372. doi:10.1016/j.chb.2010.08.015

Lane, V. R., Khuntia, J., Parthasarathy, M., & Hazarika, B. B. (2017). The Impact of the Internet on Values in India: Shifts in Self-Enhancement and Self-Transcendence Amongst Indian Youth. *Journal of Global Information Management*, *25*(3), 98–120. doi:10.4018/JGIM.2017070106

Lee, H., & Kwak, N. (2016). Mobile communication and cross-cutting discussion: A cross-national study of South Korea and the US. *Telematics and Informatics*, *33*(2), 534–545. doi:10.1016/j.tele.2015.07.006

Maglaughlin, K. L., & Sonnenwald, D. H. (2002). User perspectives on relevance criteria: A comparison among relevant, partially relevant, and not-relevant judgments. *Journal of the American Society for Information Science and Technology*, *53*(5), 327–342.

Milewski, A. E. (2007). Global and task effects in information-seeking among software engineers. *Empirical Software Engineering*, *12*(3), 311–326. doi:10.1007/s10664-007-9036-6

Nie, J.-Y. (2010). Cross-Language Information Retrieval. *Synthesis Lectures on Human Language Technologies*, *3*(1), 1–125. doi:10.2200/S00266ED1V01Y201005HLT008

Nunnally, J., & Bernstein, I. (1994). *Psychometric Theory*. New York: McGraw-Hill.

Park, T. K. (1993). The Nature of Relevance in Information Retrieval: An Empirical Study. *The Library Quarterly: Information, Community. Policy*, *63*(3), 318–351.

Roztocki, N., & Weistroffer, H. R. (2011). Information Technology Success Factors and Models in Developing and Emerging Economies. *Information Technology for Development*, *17*(3), 163–167. doi:10.1080/02681102.2011.568220

Samimi, P., Ravana, S. D., & Koh, Y. S. (2016). Effect of verbal comprehension skill and self-reported features on reliability of crowdsourced relevance judgments. *Computers in Human Behavior*, *64*, 793–804. doi:10.1016/j.chb.2016.07.058

Saracevic, T. (1975). RELEVANCE: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, *26*(6), 321–343. doi:10.1002/asi.4630260604

Saracevic, T. (2007a). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology*, *58*(13), 1915–1933. doi:10.1002/asi.20682

Saracevic, T. (2007b). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*, *58*(13), 2126–2144. doi:10.1002/asi.20681

Schwartz, A. J., Boduroglu, A., & Gutchess, A. H. (2014). Cross-Cultural Differences in Categorical Memory Errors. *Cognitive Science*, *38*(5), 997–1007. doi:10.1111/cogs.12109 PMID:24628532

Silic, M., & Back, A. (2016). What Are the Keys to a Successful Mobile Payment System? Case of Cytizi: Mobile Payment System. *Journal of Global Information Management*, *24*(3), 1–20. doi:10.4018/JGIM.2016070101

Soja, P. (2016). Reexamining Critical Success Factors for Enterprise System Adoption in Transition Economies: Learning from Polish Adopters. *Information Technology for Development*, *22*(2), 279–305. doi:10.1080/02681102.2015.1075189

Statista. (2019). Worldwide desktop market share of leading search engines from January 2010 to January 2019. Retrieved from https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/

Taylor, A. (2012). User relevance criteria choices and the information search process. *Information Processing & Management: An International Journal*, *48*(1), 136–153. doi:10.1016/j.ipm.2011.04.005

Teo, T. S. H. (2007). Organizational Characteristics, Modes of Internet Adoption and Their Impact: A Singapore Perspective. *Journal of Global Information Management*, *15*(2), 91–117. doi:10.4018/jgim.2007040104

TREC. (2019). Text Retrieval Conference. Retrieved from https://trec.nist.gov/

Vakkari, P. (1999). Task complexity, problem structure and information actions: Integrating studies on information seeking and retrieval. *Information Processing & Management: An International Journal*, *35*(6), 819–837. doi:10.1016/S0306-4573(99)00028-X

Wang, P., & Soergel, D. (1998). A cognitive model of document use during a research project. Study I. Document selection. *Journal of the American Society for Information Science*, *49*(2), 115–133. doi:10.1002/(SICI)1097-4571(199802)49:2<115::AID-ASI3>3.0.CO;2-T

Yahoo. Answers. (2019). All directories. Retrieved from https://answers.yahoo.com/dir/index

Zhang, J., & Fei, W. (2010). Search engines' responses to several search feature selections. *The International Information & Library Review*, *42*(3), 212–225. doi:10.1080/10572317.2010.10762866

Zhang, J., Fei, W., & Le, T. (2013). A comparative analysis of the search feature effectiveness of the major English and Chinese search engines. *Online Information Review*, *37*(2), 217–230. doi:10.1108/OIR-07-2011-0099

Jin Zhang is a full professor at the School of Information Studies, University of Wisconsin-Milwaukee, U.S.A. He has published papers in journals such as Journal of the American Society for Information Science and Technology, Information Processing & Management, Journal of Documentation, Journal of Intelligent Information Systems, Online Information Review, etc. His book "Visualization for Information Retrieval" was published in the Information Retrieval Series by Springer in 2008. His research interests include visualization for information retrieval, information retrieval algorithm, metadata, search engine evaluation, consumer health informatics, social media, digital libraries, data mining, knowledge system evaluation, and human computer interface design.

Yuehua Zhao is a research assistant professor at the School of Information Management, Nanjing University, China. Her research interests span a variety of areas including data science, data analytics, text analysis, consumer health informatics, social media research, social network analysis, information visualization, informetrics, and scholarly communication.

Xin Cai is a Ph.D. candidate in the iSchool at the University of Wisconsin-Milwaukee. He holds a master's degree in information science from Central China Normal University, China, and a bachelor's degree in computer science from Shenyang Normal University, China. His research interests include information retrieval and systems, data mining, and domain analysis.

Taowen Le, tenured professor and former department chair of information systems & technologies at Weber State University, Business Division Chair of Utah Academy of Science, Arts, and Letter.

Feicheng Ma is a Professor at the School of Information Management, WuHan University. His research interests include social informatization, web information governance, information resources planning and management, and big data analysis.